Statistical Software in the Higher School Educational Process

Taras Kobylnyk

Drohobych Ivan Franko State Pedagogical University, 24 Ivan Franko Str., Drohobych, Ukraine
kobylnyktaras@gmail.com

Abstract. Data processing is not possible without a computer with the appropriate software. Before the user there is a problem of choosing software for research. Therefore the aim was to analyze software for statistical research. To achieve the goal the scientific and methodological sources were analyzed, comparison, generalization were made to determine the state of a problem and perspective directions of its solution. The paper presents the software capabilities of such tools as MS Excel, Statistica, IBM SPSS, R. Special attention is focused on the calling R from other software, including Statistica, IBM SPSS, Mathematica, LaTeX. The popular commercial statistical software provides the ability to perform R script directly in the shell of these programs. It is proposed to use R package for the scientific research and in the higher school educational process. In the educational process it is appropriate to use the project method. Students made report and presentation by using integrating R into LaTeX.

Keywords: Statistical Analysis**,** MS Excel, R, IBM SPSS, Statistica, Mathematica, LaTeX.

Key Terms. SoftwareSystem, ICTEnvironment, TeachingProcess

1 Introduction

In the experimental research the different statistical methods were used to test hypotheses, constructing statistical models of objects, phenomena, regularities and processes. It is obviously that the experimental data processing is impossible without the use of computers with appropriate software. There is a large variety of software including general, special purpose for data processing. Thus, the user raises the question: which software to choose? Or which software is best for data processing?

Standard statistical methods of data processing were implemented in spreadsheets (Lotus, QuatroPro MS Excel, OpenOffice.org Calc and other), Computer Mathematics Systems (CMS) (Maple, Mathematica, Matlab, Maxima and other), specialized software (R, SPSS, Statistica, SAS and other).

Often, in the learning process of higher school the students study and use specialized software such as MS Excel, IBM SPSS, Statistica for statistical analysis. The book [10] provides a comprehensive coverage of the main statistical analysis topics (data description, statistical inference, classification and regression, factor analysis, survival data, directional statistics) that one faces in practical problems, discussing their solutions with Statistica, SPSS, Matlab and R. The book [4] provides a brief and straightforward description of how to conduct a range of statistical analyses using of SAS, version 9.2. The book [18] covers the applying of many of the most powerful statistical and machine learning techniques used in data science projects in using the programming language R to perform actual research data processing. The article [12] presents various ways of measuring of the popularity or market share of software for advanced analytics. Such software is also referred to as tools for data science, statistical analysis, machine learning, artificial intelligence, predictive analytics, business analytics, and is also a subset of business intelligence.

Research methods. Scientific and methodological sources were analyzed, comparison, generalization was used to determine the state of the problem and perspective directions of its solution.

2 The Presentation of Main Results

2.1 Microsoft Excel.

The book [5] shows how to use Microsoft Excel to perform statistical analysis. This step-by-step guide has been updated to cover the new features and interface of Excel 2010. There are advantages of Microsoft Excel:

- the relative ease of mastering and practical use;
- a significant number of built-in statistical functions;
- the presence of the add "Analysis ToolPak", containing procedures for solving complex problems of statistical analysis;
- the ability to create user "own" modules in the VBA for data analysis;
- there is a macro-addition xlstat-pro (You can use the trial from site developer www.xlstat.com).

Despite the advantages of MS Excel a full statistical analysis is not possible: this is software of general purpose but not special (scientific). There is a small set of statistical tests. There are not many methods, especially multidimensional. There is no specialized reporting system. This is explained by the fact that the statistics is not the primary function of a spreadsheet.

Statistical packages are deprived of these deficiencies, far surpassing spreadsheets by the volume and quality of implemented statistical methods. Therefore, the final statistical analysis should be done in the software that is specifically designed for this purpose - statistical packages. The statistical package usually contains quite a large range of standard statistical methods, simple enough for rapid studying, working with rather large databases, it is possible to exchange data with other packages and databases, it has an extensive set of graphics data and the results of their analysis and there is a detailed documentary support and reference system [16].

Statistical package must satisfy the following minimum requirements:

- modularity;
- use a simple problem-oriented language for the formulation of user tasks;
- maintaining the user's data bank and report on the results of this analysis;
- interactive user mode with the package;
- compatibility from other software.

It should be noted there is a minimal set of statistical methods of analysis, which is included in all statistical packages, such as descriptive statistics, nonparametric statistics, variance, correlation and regression, cluster, factor, discriminant analyzes.

2.2 IBM SPSS

IBM SPSS (Statistical Package for Social Science) is a universal statistical package of the company SPSS Inc. The first version of the package was released in 1968. In 2009, IBM swallowed SPSS Inc, whereby the name of the package includes the IBM abbreviation. IBM SPSS is modular software. Its foundation is the base module (SPSS Base), allowing to carry out all types of statistical analysis. The latest version is 24. The book [7] has been written for use with version 20 of the IBM SPSS statistical package for Windows. This book focuses on both the Windows method and the syntax method of analysis, which has proven popular with both experienced researchers and students.

IBM SPSS is among the most widely used software for statistical analysis in social science and other disciplines. IBM SPSS can take data from almost any type of file and use them to generate tabulated reports, charts and plots of distributions and trends, descriptive statistics, and complex statistical analyses.

Add-on modules can be installed for the extended and in-depth data analysis, such as IBM SPSS Missing Values (finds relationships between any missing values in your data and other variables), IBM SPSS Neural Networks (offers non-linear data modeling procedures that enable you to discover more complex relationships in your data), IBM SPSS Forecasting (enables analysts to predict trends and develop forecasts quickly and easily – without being an expert statistician) and other. More detailed information about the modules can be found at the site http://www-03.ibm.com/software/products/en/spss-statistics.

Among the advantages of the IBM SPSS should be noted:

- a simple and user-friendly interface;
- a wide variety of statistical and graphical data analysis procedures and reporting procedures;
- the detailed context-based help system;
- the ability to download trial version on the official website of the company, the existence software versions in many languages;
- compatibility with operating systems Windows, Mac OS, Linux.

Among the disadvantages of the IBM SPSS package should be noted:

- high system requirements (requires 1GB or more of RAM, 800 MB hard disk memory and a CPU frequency of 1 GHz and above);
- high price compared to similar software.

2.3 Statistica

Statistica [17] is universal software for statistical analysis of StatSoft Inc. The first version of the package (Statistica for DOS) was released in 1991. To date, it developed 13-th version of the package.

There are three standard components (modules) of Statistica:

- Statistica Base module provides an extensive choice of the main types of statistical analysis. It takes requires 256 MB RAM or more and CPU frequency of 500 MHz and above for efficient operation of the Statistica Base module;
- Advanced Linear/NonLinearModels module is used for modeling and forecasting, including the ability to automatically select the model and advanced interactive visualization tools;
- Multivariate Exploratory Techniques module is used for exploratory analysis of the different types of data in combination with interactive visualization tools.

There are such options for installing Statistica:

- a single-user version (Single-User);
- a network version (Concurrent Network) - version for use in local area networks;
- Enterprise – version for use in computer systems and large organizations;
- Web-Based – version for use in large networks through a browser.

Among the advantages of Statistica should be noted:

- there is implemented communication between Statistica and Windows-applications;
- there is a built-in programming Statistica Basic language.

The disadvantages include the fact that Statistica is developed only for Windows, which slightly decrease the number of users and the high cost of licenses. Statistica does not have a version for Mac OS or Linux. Statistica relies on Microsoft's .NET Framework and other Microsoft technologies. But thanks to advancement in computer hardware and software, Statistica can run on a Windows OS inside of a virtual machine on a Mac or Linux computer. Some examples of virtualization software for the Mac include VMware Fusion and Parallels. More information on this topic can be found on the site https://support.quest.com/statistica/kb/145707.

2.4 R Environment

First of all R is a high-level language and an environment for data analysis and graphics [2]. Learning of statistical analyses with R and global scientific community support has resulted that R-scripts are the standard for both of articles and in informal communication of scientists around the world.

We list the advantages and disadvantages R package:

- it is a free software, somebody can download it from http://www.r-project.org;
- there is a realization of the operating systems Windows, Mac OS, Linux;
- R basic version takes up little space on hard disk and includes all the functions needed for statistical analysis;
- package extensions is developed that apply in virtually all fields of science where statistical analysis is used;
- there is an opportunity to write the necessary functions by himself;
- data can be entered from the keyboard or import from text files, spreadsheets, statistical software, CMS, database management systems.

Among the disadvantages of conventionally should be noted:

- orientation to programming;
- in contrast to most commercial software R has not several Graphical User Interfaces, but has Command Line Interface and thus need to know the necessary functions and programming language syntax. It should be noted there are several Graphical User Interfaces for R (RStudio, R Commander and others), but they are not as good as the interfaces.

2.5 Computer Mathematics Systems

If statistical analysis is only part of some large-scale scientific research in this case you can use CMS, for example Maple [14], Mathematica [15] or Matlab [11]. CMS covers virtually all areas of classical and Applied Mathematics and Statistics. Normally some basic statistics functions are in the CMS kernel. Package development (or Toolbox) is used to increase the possibilities of using to data processing.

Maple, Mathematica, Matlab are commercial and costly CMS. The Maxima, a Computer Algebra System, is the real alternative to commercial CMS (http://maxima.sourceforge.net/). The Maxima source code can be compiled on many systems, including Windows, Linux and Mac OS. In fact, it is the only system that can compete with commercial Mathematica and Maple, but it has several limited possibilities for statistical analysis.

3 Calling R from Other Software

Today there is a trend of integration of statistical packages between themselves and other software, including CMS and word processors. We will consider examples of such integration. It should be noted that most achievements in statistics are realized for the first time in package R and then added to the menu other software. The R software is powerful but it takes a long time to learn to use it well. You can keep using your current software to access and manage data, then call R for just the things your current software doesn’t do [13]. Therefore, we consider the integrating R into other software such as Statistica, IBM SPSS, Mathematica and word processors.

3.1 Calling R from Statistica

According to information at the [8] R tabular results are difficult to handle; R graphics cannot be changed; results difficult to manage. Statistica and R integration expands the possibility of data environments, open and use R scripts inside Statistica. By integrating R into Statistica, the user can run R scripts in the shell Statistica, obtaining results in the form of reports, workbooks and charts Statistica. You can also call methods R from Statistica Visual Basic. Statistica interacts with other applications using the relevant standards, in particular by using the COM standard, which is built into the Windows. Statistica interacts with R through statconnDCOM and, if the library is installed in the system, users can open and execute R scripts from Statistica environment. It facilitates the transfer of data and presentation of results through custom macro running in Statistica, instead of R. Interaction with R is available in all Statistica product lines. From the 10th version of Statistica, the process of establishing interaction with R is carried out automatically.

3.2 Calling R from IBM SPSS

According to the information at the [3] the 16th version of IBM SPSS package contains a free plug-in that allows you to execute code R in IBM SPSS. With the R plug-in, you retain all the features of an SPSS database, particularly the labels of category data and the long descriptors. The R integration plug-in does two things:

- it opens communication between IBM SPSS and R;
- it provides R with a package of functions with which to translate IBM SPSS data structures into R objects.

More about integrating R into IBM SPSS you can read in the article [3]. Thus, the functionality of R is available for users who do not have the skills in working with R.

3.3 Calling R from Mathematica

By integrating the R code in Mathematica data is exchanged between Mathematica and R, and the execution of the R code in Mathematica. From version 9, Mathematica offers built-in ways to integrate R code into your Mathematica workflow, combining Mathematica's broad range of capabilities with the R. There are built-in ways to integrate R code into the Mathematica system using the RLink package development. RLink uses J/Link and rJava/JRI Java library to help users exchange data between Mathematica and R, and in order to execute code P from Mathematica.

Calling R from Mathematica provides:

- full access to all R functionality from within Mathematica;

- exchange data between Mathematica and R, using the usual Mathematica representation for the data – most core R data types are supported;
- apply built-in and user-defined R functions from within Mathematica in a way natural for Mathematica users;
- create functions in the R workspace from within Mathematica.

More about calling R from Mathematica you can read in the article [1].

3.4 Calling R from Word Processors

Usually, any research is completed publication of results. To save the results suitable for publication as is usually used two packages – sweave and odfWeave from R. The first allows you to embed the code and results in the tex-files for typographic quality reports in the formats pdf, PostScript and dvi. Similarly odfWeave package allows you to embed the code and the results of R in the file format odf (Open Documents Format). The final report, you can edit at word processor OpenOffice Writer or Microsoft Word. More calling R from word processors you can read in the book [9].

4 Conclusions and Prospects for the Further Research

There is question about software choice for data processing for users. Of course, functionality, high quality and reasonable price are the factors for choosing statistical software. Therefore, you need to consider when choosing statistical software:

- the cost (if system is commercial);
- the purpose of use (for scientific research or educational process);
- the user qualifications requirements (level of statistical knowledge and programming);
- hardware and software available.

From the above software analysis we advise to use R package because:

- it is distributed under license GNU/GPL;
- it is one of the best statistical software.

The last argument is confirmed by the fact that the popular commercial statistical systems (IBM SPSS, Statistica, SAS), CMS (Mathematica, Matlab) provides the ability to perform R script directly in the shell of these programs. Integrating R into other software can be useful not only to specialists from the statistics, but also specialists, for whom data processing is only part of the big scientific research.

In the educational process it is appropriate to use the project method. Students are encouraged to implement the project; the end result is to report on the implementation of individual tasks. The subject is related data processing by using of R package. Report and presentation are made by using integrating R into LaTeX.

The use of statistical software extends the scope of application of mathematical methods and models for the study of processes in various fields of human activity and simplifies them. The final result of research is some publication. In this case LaTeX is useful which combined with the R package or CMS provides the ability to create high quality products. Further research will be focused to exploring the possibilities of using R environment for various statistical analysis methods and development of methods of its use in the higher school educational process.

References

1. Built-in Integration with R: New in Mathematica 9, https://www.wolfram.com/mathematica/new-in-9/built-in-integration-with-r/
2. Crawley, M.J.: The R Book, Second Edition. A John Wiley & Sons, Chichester (2013)
3. Dalzell, C.: Calling R from SPSS, https://www.ibm.com/developerworks/library/ba-call-r-spss/
4. Der, G., Everitt, B.S.: A Handbook of Statistical Analyses Using SAS, Third Edition. Chapman and Hall/CRC Press, Boca Raton (2008)
5. Dretzke, B.J. Statistics with Microsoft Excel, 5th Edition. Pearson, New Jersey (2011).
6. Field, A.: Discovering Statistics Using IBM SPSS Statistics, 4th Edition. Sage Publications, London (2013)
7. Ho, R.: Handbook of Univariate and Multivariate Data Analysis with IBM SPSS, Second Edition. Chapman and Hall/CRC, Boca Raton (2013)
8. Integrating R into Statistica, http://statsoft.ru/products/integration/integration-with-R.php (in Russian)
9. Kabacoff, R.I.: R in Action: Data Analysis and Graphics with R, Second Edition. Manning Publications, NY (2015)
10. Marques de Sá, Joaquim P.: Applied Statistics Using SPSS, STATISTICA, MATLAB and R, Second Edition. Springer Publishing Company, Incorporated, Berlin (2007)
11. Martinez, W.L., Martinez, A.R.: Exploratory Data Analysis with Matlab. Chapman & Hall/CRC Press, Boca Raton (2005)
12. Muenchen, R.A.: The Popularity of Data Science Software, http://r4stats.com/articles/popularity
13. Muenchen, R. A. Calling R from Other Software, http://r4stats.com/articles/calling-r/
14. Rafter, J.A., Abell, M.L., James P.B. Statistics with Maple. Academic Press, Amsterdam; Boston (2002)
15. Rose, C., Smith, Murray D. Mathematical Statistics with MATHEMATICA. Springer-Verlag, NY (2002)
16. Tyurin, Yu. N., Makarova, A.A. Statistical analysis of the data on the computer. INFRA, Moscow (1998) (in Russian)
17. Wass, J.A. STATISTICA 10: The Power of Statistics and Data Mining Simplified, http://www.scientificcomputing.com/article/2011/12/statistica-10-power-statistics-and-data-mining-simplified
18. Zumel, N., Mount, J.: Practical Data Science with R. Manning Publications, NY (2014)