Zhurtbay N.T., Mamyrova G.S.

The role of statistical programs in research

Statistical software are specialized computer programs for analysis in statistics. As quantitative research grows, application of statistical software (SS) becomes a more crucial part of data analysis. Researchers are experiencing a transition from manual analysis with paper to more efficient digital/electronic analysis with statistical software (SS). In this article the authors examine the program in terms of efficiency of use in the humanities.

Key words: statistic, computer programs, research, quantitative research.

Журтбай Н.Т., Мамырова Г.С.

Статистикалық бағдарламалардың ғылыми зерттеулердегі рөлі

Статистикалық бағдарламалар – статистика саласындағы саралауға арналған маманданған компьютерлік бағдарламалар. Сандық зерттеулердің қолдану аясының ұлғаюына байланысты статистикалық бағдарламалық қамту (statistic software) саралаудың маңызды бөлігіне айналып келеді. Зерттеушілер қолмен орындалатын қағаз зерттеулерден тиімдірек келетін цифрлі/электронды саралау статистикалық бағдарламалық қамтуға (statistic software) көшу кезеңінде. Берілген мақалада авторлар бұл бағдарламалардың гуманитарлық ғылымдарда қолданылу тиімділігін қарастырады.

Түйін сөздер: статистика, компьютерлік бағдарламалар, ғылыми-зерттеу, ғылыми сандық зерттеу.

Журтбай Н.Т., Мамырова Г.С.

Роль статистических программ в области научных исследований

Статистические программы – специализированные компьютерные программы для анализа в области статистики. По мере роста количественных исследований, применение статистического программного обеспечения (Statistical Software) становится все более важной частью анализа данных. Исследователи переживают переход от ручного анализа с бумагой для более эффективного цифрового / электронного анализа с помощью статистического программного обеспечения (Statistical Software). В данной статье авторы рассматривают программы с точки зрения эффективности использования в гуманитарных науках.

Ключевые слова: статистика, компьютерные программы, исследования, количественные исследования.

UDC 070:502(574)

Zhurtbay N.T.[1], Mamyrova G.S.[2],

[1]PhD, Senior Lecturer, [2]Senior Lecturer, of KazNU named
after Al–Farabi, Almaty, Kazakhstan,
e-mail: nazym–zhurtbay@mail.ru; mgs1801@mail.ru

THE ROLE OF STATISTICAL PROGRAMS IN RESEARCH

Introduction

Statistical programs are becoming increasingly important in the work of researchers in Kazakhstan. The researchers in their scientific work shows not only in theory but also solutions numerals indicators. And on the basis of indicators of numerals appear visually informative info graphics. In this article, we will conduct a comparative analysis of statistical programs that will be useful in the work of scientists from different field. The main issue of this study is to determine the most suitable software for research in the humanitarian area?

The study will be considered free, partly free and paid statistical software. Website http://statpages.info/ classifies statistical programs by follows:

Table 1

1	General Packages:	support a wide variety of statistical analyses
2	Subset Packages:	deal with a specific area of analysis, or a limited set of tests
3	Curve Fitting and Modeling:	to handle complex, nonlinear models and systems
4	Biostatistics and Epidemiology	especially useful in the life sciences
5	Surveys, Testing and Measurement:	especially useful in the business and social sciences
6	Excel Spreadsheets and Add–ins:	you need a recent version of Excel
7	Programming Languages and Subroutine Libraries: customized for statistical calculations	you need to learn the appropriate syntax
8	Scripts and Macros	for scriptable packages, like SAS, SPSS, R, etc
9	Miscellanious	don’t fit into any of the other categories

In the main part we examine the statistical program with the first group of General packages (Table 1). This next program: OpenStat, JASP, Develve, Explorer, SalStat–2, SOFA, ViSta, PSPP, OpenEpi Version 2.3, Statext, MicrOsiris, Gnumeric, Statist, Tanagra, Dap, PAST, AM, Instat Plus, WinIDAMS, SSP (Smith's Statistical Package), Dataplot, WebStat, Regress+, SISA, Statistical Software, IRRISTAT, Data Desk, MaxStat, SHAZAM, SYSTAT 12, Statlets, Wizard, WINKS, StudyResult, STATGRAPHICS Plus v5.0, NCSS–2007 (Statistical Analysis System), PASS–2008 (Power and Sample Size, and GESS (Gene Expression software for Micro–arrays), MiniTab, InStat, Prism, CoStat 6.2, AppOnFly [2].

Main body

Completely Free. Can be freely downloaded and used in their fully–functional mode (no strings attached). OpenStat a general stats package for all Windows versions (including Win 7 and Win 8) and for Linux systems (under Wine), developed by Bill Miller of Iowa State U, with a very broad range of data manipulation and analysis capabilities and an SPSS–like user interface. Bill also has provided an excellent User Manual as an Adobe Acrobat file. In addition, there is a free Pascal program, manual, sample data and source code for LazStats which contains programs similar to OpenStat [1].

SPSS Statistics is a not free software package used for statistical analysis. Long produced by SPSS Inc., it was acquired by IBM in 2009 [3]. The current versions (2015) are officially named IBM SPSS Statistics. Companion products in the same family are used for survey authoring and deployment (IBM SPSS Data Collection), data mining (IBM SPSS Modeler), text analytics, and collaboration and deployment (batch and automated scoring services) [8].

The software name originally stood for Statistical Package for the Social Sciences (SPSS), reflecting the original market, although the software is now popular in other fields as well, including the health sciences and marketing [5].

SPSS is a widely used program for statistical analysis in social science. It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners, and others. The original SPSS manual (Nie, Bent & Hull, 1970) has been described as one of «sociology's most influential books» for allowing ordinary researchers to do their own statistical analysis. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary was stored in the datafile) are features of the base software [9].

Statistics included in the base software:

– Descriptive statistics: Cross tabulation, Frequencies, Descriptives, Explore, Descriptive Ratio Statistics

– Bivariate statistics: Means, t–test, ANOVA, Correlation (bivariate, partial, distances), Nonparametric tests

– Prediction for numerical outcomes: Linear regression

– Prediction for identifying groups: Factor analysis, cluster analysis (two–step, K–means, hierarchical), Discriminant

Free, but «demonstration» or «student versions» of commercial packages; can be freely downloaded, but are usually restricted or limited in some way.

JASP – a new package (still in «beta» development) that's described by the authors as a «low–fat alternative to SPSS», and «Bayesian statistics made accessible». Provides a user–friendly interface to many of the commonly–used statistical analyses – descriptive statistics, plots, t tests, Levene's Test, ANOVA, ANCOVA, contingency tables, Pearson and Spearman correlation, Kendall's Tau–B, and linear regression [10]. For many of these analyses, the JASP also provides the closest corresponding Bayesian equivalent, implemented in a way that will be understandable to people not familiar with Bayesian concepts and terminology. Develve – stats package for fast and easy interpretation of experimental data in science and R&D in a technical environment. Everything is directly accessible and results are directly visible, with no hidden menus; e.g.: graphs are easily scrollable, and when clicked, a bigger version pops up. Results for group comparisons directly indicate the significance of the difference in average and variation, and if the sample size is sufficiently large. Has a basic mode for statistical testing, and a design–of–experiments mode. Explorer – A data exploration / graphing / analysis program with a very elegant drag–and–drop interface. Accepts data from text files, Excel spreadsheet, MySQL databases, and copy/pasted from the clipboard. Provides over a dozen kinds of plots and diagrams, basic statistical summaries, significance tests (chi–square, t, ANOVA) and more advanced analyses (factorial, principal components, discriminant, variance, linear regression). Executable programs can be downloaded for Windows and Mac OSX. Written in JavaScript, so it can also be run in any modern browser.

SalStat–2 – a multi–platform, easy–to–use statistical system that provides data management (importing, editing, pivot tables), statistical calculations (descriptive summaries, probability functions, chi–square, t–tests, 1–way ANOVA, regression, correlation, non–parametric tests, Six–Sigma), and graphs (bar, line, scatter, area, histogram, box&whisker, stem, adaptive, ternary scatter, normal probability, quality control).

SOFA (Statistics Open For All) – an innovative statistics, analysis, and reporting program. Available for Windows, Mac and Linux systems. Has an emphasis on ease of use, learn as you go, and beautiful output.

ViSta – a Visual Statistics program for Win3.1, Win 95/NT, Mac and Unix, featuring a Structured Desktop, with features designed to structure and assist the statistical analyst.

PSPP – a free replacement for SPSS (although at this time it implements only a small fraction of SPSS's analyses). But it's free, and will never «expire». It replicates the «look and feel» of SPSS very closely, and even reads native SPSS syntax and files [11].

– Supports over 1 billion cases and over 1 billion variables.

– Choice of terminal or graphical user interface; Choice of text, postscript or html output formats.

– Inter–operates with Gnumeric, OpenOffice. Org and other free software.

– Easy data import from spreadsheets, text files and database sources.

– Fast statistical procedures, even on very large data sets.

– No license fees; no expiration period; no unethical «end user license agreements».

– Fully indexed user manual.

– Cross platform; Runs on many different computers and many different operating systems. Note: For Windows installer.

OpenEpi Version 2.3 – OpenEpi is a free, web–based, open source, operating–system–independent series of programs for use in public health and medicine, providing a number of epidemiologic and statistical tools. Version 2 (4/25/2007) has a new interface that presents results without using pop–up windows, and has better installation methods so that it can be run without an internet connection. Version 2.2 (2007/11/09) lets users run the software in English, French, Spanish, or Italian.

Statext – Provides a nice assortment of basic statistical tests, with text output (and text–based graphics). Capabilities include: rearrange, transpose, tabulate and count data; random sample; basic descriptives; text–plots for dot, box–and–whiskers, stem–and–leaf, histogram, scatterplot; find z–values, confidence interval for means, t–tests (one and two group, and paired; one– and two–way ANOVA; Pearson, Spearman and Kendall correlation; ;inear regression, Chi–square goodness–of–fit test and independence tests; sign test, Mann–Whitney U and Kruskal–Wallis H tests, probability tables (z, t, Chi–square, F, U); random number generator; Central Limit Theorem, Chi–square distribution [12].

MicrOsiris – a comprehensive statistical and data management package for Windows, derived from the OSIRIS IV package developed at the University of Michigan. It was developed for serious survey analysis using moderate to large data sets. Main features: handles any size data set; has Excel data entry; imports/exports SPSS, SAS, and Stats datasets; reads ICPSR (OSIRIS) and UNESCO (IDAMS) datasets; data mining techniques for market analysis (SEARCH – very fast for large datasets); interactive decision tree for selecting appropriate tests; database maniuplation (dictionaries, sorting, merging, consistency checking, recoding, transforming) extensive statistics (univariate, staccerplot, cross–tabs, ANOVA/MANOVA, log–linear, correlation/regressionMCA, MNA, binary segmentation, cluster, factor, MINISSA, item analysis, survival analysis, internal consistency); online, web–enabled user's manual; requires only 6MB RAM; uses 12MB disk, including manual. Fully–functional version is free; the authors would appreciate a small donation to support ongoing development and distribution.

Gnumeric – a high–powered spreadsheet with better statistical features than Excel. Has 60 extra functions, basic support for financial derivatives (Black Scholes) and telecommunication engineering, advanced statistical analysis, extensive random number generation, linear and non–linear solvers, implicit intersection, implicit iteration, goal seek, and Monte Carlo simulation tools [14, 24].

Statist – a compact, portable program that provides most basic statistical capabilities: data manipulation (recoding, transforming, selecting), descriptive stats (including histograms, box&whisker plots), correlation & regression, and the common significance tests (chi–square, t–test, etc.). Written in C (source available); runs on Unix/Linux, Windows, Mac, among others.

Tanagra – a free (open–source) data–mining package, which supports the standard «stream diagram» paradigm used by most data–mining systems. Contains components for Data source (tab–delimited text), Visualization (grid, scatterplots), Descriptive statistics (cross–tab, ANOVA, correlation), Instance selection (sampling, stratified), Feature selection and construction, Regression (multiple linear), Factorial analysis (principal components, multiple correspondence), Clustering (kMeans, SOM, LVQ, HAC), Supervised learning (logistic regr., k–NN, multi–layer perceptron, prototype–NN, ID3, discriminant analysis, naive Bayes, radial basis function), Meta–spv learning (instance Spv, arcing, boosting, bagging), Learning assessment (train–test, cross–validation), and Association (Agrawal a–priori).

Dap – a statistics and graphics package developed by Susan Bassein for Unix and Linux systems, with commonly–needed data management, analysis, and graphics (univariate statistics, correlations and regression, ANOVA, categorical data analysis, logistic regression, and nonparametric analyses). Provides some of the core functionality of SAS, and is able to read and run many (but not all) SAS program files. Dap is freely distributed under a GNU–style «copyleft»[13].

PAST – an easy–to–use data analysis package aimed at paleontology including a large selection of common statistical, plotting and modelling functions: a spreadsheet–type data entry form, graphing, curve fitting, significance tests (F, t, permutation t, Chi–squared w. permutation test, Kolmogorov–Smirnov, Mann–Whitney, Shapiro–Wilk, Spearman's Rho and Kendall's Tau tests, correlation, covariance, contingency tables, one–way ANOVA, Kruskal–Wallis test), diversity and similarity indices & profiles, abundance model fitting, multivariate statistics, time series analysis, geometrical analysis, parsimony analysis (cladistics), and biostratigraphy.

AM – a free package for analyzing data from complex samples, especially large–scale assessments, as well as non–assessment survey data. Has sophisticated stats, easy drag & drop interface, and integrated help system that explains the statistics as well as how to use the system. Can estimate models via marginal maximum likelihood (MML), which defines a probability distribution over the proficiency scale. Also analyzes «plausible values» used in programs like NAEP. Automatically provides appropriate standard errors for complex samples via Taylor–series approximation, jackknife & other replication techniques.

Instat Plus – from the University of Reading, in the UK. (Not to be confused with Instat from GraphPad Software.) An interactive statistics package for Windows or DOS.

WinIDAMS – from UNESCO – for numerical information processing and statistical analysis. Provides data manipulation and validation facilities classical and advanced statistical techniques, including interactive construction of multidimensional tables, graphical exploration of data (3D scattergram spinning, etc.), time series analysis, and a large number of multivariate techniques.

SSP (Smith's Statistical Package) – a simple, user–friendly package for Mac and Windows that can enter/edit/transform/import/export data, calculate basic summaries, prepare charts, evaluate distribution function probabilities, perform simulations, compare means & proportions, do ANOVA's, Chi Square tests, simple & multiple regressions.

Dataplot – (Unix, Linux, PC–DOS, Windows) for scientific visualization, statistical analysis, and non–linear modeling. Has extensive mathematical and graphical capabilities. Closely integrated with the NIST/SEMATECH Engineering Statistics Handbook.

WebStat – A Java–based statistical computing environment for the World Wide Web. Needs a browser, but can be downloaded and run offline.

Regress+ – A professional package (Macintosh only) for univariate mathematical modeling (equations and distributions). The most powerful software of its kind available anywhere, with state–of–the–art functionality and user–friendliness. Too many features to even begin to list here.

SISA – Simple Interactive Statistical Analysis for PC (DOS) from Daan Uitenbroek. An excellent collection of individual DOS modules for several statistical calculations, including some analyses not readily available elsewhere.

Statistical Software by Paul W. Mielke Jr. – a large collection of executable DOS programs (and Fortran source). Includes: Matrix occupancy, exact g–sample empirical coverage test, interactions of exact analyses, spectral decomposition analysis, exact mrbp (randomized block) analyses, exact multi–response permutation procedure, Fisher's Exact for cross–classfication and goodness–of–fit, Fisher's combined p–values (meta analysis), largest part's proportion, Pearson–Zelterman, Greenwood–Moran and Kendall–Sherman goodness–of–fit, runs tests, multivariate Hotelling's test, least–absolute–deviation regression, sequential permutation procedures, LAD regression, principal component analysis, matched pair permutation, r by c contingency tables, r–way contingency tables, and Jonkheere–Terpstra.

IRRISTAT – for data management and basic statistical analysis of experimental data (Windows). Primarily for analysis of agricultural field trials, but many features can be used for analysis of data from other sources. Includes: Data management with a spreadsheet , Text editor, Analysis of variance, Regression, Genotype x environment interaction analysis, Quantitative trait analysis, Single site analysis, Pattern analysis, Graphics, Utilities for randomization and layout, general factorial EMS, and orthogonal polynomial.

Data Desk – first released in 1986, is one of the oldest Mac programs still actively developed. The modern versions (for Mac OS X and Windows computers) are available for sale from Data Description, Inc. But in honor of its origin, there is also a free version that runs on Macintosh 680x0 computers (which were made from 1984 to 1996). While this old hardware is now ancient history, there are software emulators of 680x0 computers that run on modern Mac, Windows and Linux computers (see the Gryphel Project) [15].

Conclusion

We reviewed 42 statistical programs. The following conclusions were made based on the information accompanying the developers to statistical programs.

Free programs such as OpenStat very limited research capabilities. And requires experience with statistical programs.

Partially free software in free mode functioning partially. In demo version you can use once, but if the program is suitable for your particular study is a working version you will need to buy and install on your laptop. This is very useful in order to avoid the purchase of unsuitable programs.

References

1 S. Matthew Sunday Abatan, Michael Sunday Olayemi. The Role of Statistical Software in Data Analysis. International Journal of Applied Research and Studies (iJARS) ISSN: 2278-9480 Volume 3, Issue 8 (August – 2014).

2 A. Alexeyev, O.G. Echevskaya, G.D. Kovaleva, P.S. Rostovtsev. Analysis of sociological data using the SPSS package. Collection of case studies. – Novosibirsk: Publishing center of the NSU 2003.

3 Byuyul Achim, Tsëfel Peter. SPSS: data processing Art. Analysis of statistical data and restore hidden patterns: Trans. with it. / Achim Byuyul Peter Tsëfel – Spb .: «DiaSoftYuP» 2005-608 p.

4 Nasledov AD Mathematical methods of psychological research: Analysis and interpretation of data: Textbook. – SPb .: Rech, 2004. – 392 p.

5 http://statpages.info

6 Nasledov AD SPSS: Computer data analysis in psychology and social sciences. 2nd ed. – SPb.: Peter, 2006 ISBN 978-5-91180-318-6

7 Patsiorkovskii VV VV Patsiorkovskii SPSS for social scientists. Tutorial. – M .: ISESP RAS, 2005. – 433 p.

8 Andy Field. Discovering Statistics Using SPSS, Second Edition. – 2005.

9 Griffith A. SPSS for Dummies. – Hoboken: Wiley Publishing, 2007.

10 Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2004). SPSS for introductory statistics: Use and Interpretation. Mahwah, NJ: Lawrence Erlbaum Associates.

11 Leech, N. A., Barett K.C., & Morgan, G.A. (2004). SPSS for intermediate statistics: Use and interpretation.

12 Gustav Levine, Sanford L. Braver, David P. Mackinnon, Melanie C. Page, Gustav. Levine's Guide to SPSS for Analysis of Variance. – 2nd ed. – Lawrence Erlbaum Associates; 2nd edition, 2005. – 200 p. – ISBN 978-0805830958.

13 Programming and Data Management for SPSS 16.0: A Guide for SPSS and SAS Users. – ISBN 978-1-56827-399-0

14 SPSS. (2003). SPSS 12.0: Brief guide. Chicago: Author.

15 Vijay Gupta. SPSS for Beginners. 1999.