COMPARISON OF COMMONLY USED STATISTICS PACKAGE PROGRAMS

Ömer Faruk KARAOKUR[1]*, Fahrettin KAYA[2], Esra YAVUZ[1], Aysel YENİPINAR[1]

[1]*Kahramanmaraş Sütçü Imam University, Faculty of Agriculture, Department of Animal Science, 46040, Onikişubat, Kahramanmaraş, Turkey*

[2]*Kahramanmaraş Sütçü Imam University, Andırın Vocational School, Computer Technology Department, 46410, Andırın, Kahramanmaras, Turkey.*

Abstract

The specialized computer programs used in the collection, organization, analysis, interpretation and presentation of the data are known as statistical software. Descriptive statistics and inferential statistics are two main statistical methodologies in some of the software used in data analysis. Descriptive statistics summarizes data in a sample using indices such as mean or standard deviation. Inferential statistics draws conclusions that are subject to random variables such as observational errors and sampling variation. In this study, statistical software used in data analysis is examined under two main headings as open source (free) and licensed (paid). For this purpose, 5 most commonly used software were selected from each groups. Statistical analyzes and analysis outputs of this selected software have been examined comparatively. As a result of this study, the features of licensed and unlicensed programs are presented to the researchers in a comparative way.

Keywords: Statistical software, Statistical analysis, Data analysis

*Corresponding author: Kahramanmaraş Sütçü Imam University, Faculty of Agriculture, Department of Animal Science, 46040, Onikişubat, Kahramanmaraş, Turkey
E mail: karaokur@gmail.com (Ö. F. KARAOKUR)
Ömer Faruk KARAOKUR   https://orcid.org/ 0000-0002-3436-8415
Fahrettin KAYA   https://orcid.org/ 0000-0003-1666-4859
Esra YAVUZ   https://orcid.org/ 0000-0002-5589-297X
Aysel YENİPINAR   https://orcid.org/ 0000-0001-2345-6789

1. Introduction

Statistical analysis process is the collection, organization, analysis, interpretation and presentation of data. It starts with defining the method and population. The population consists of observations collected by selected methods at various times, and the data obtained from each observation serves as a member of the overall group.

Statistical analysis process examines in two groups, which are descriptive statistics and inferential statistics.

Descriptive statistics summarize population data by describing what was observed in the sample graphically or numerically. The numeric descriptors are the mean and standard deviation for continuous data types, and the percent and frequency for categorical data.

Inferential statistics allows making inferences about population using patterns in the sample by taking into consideration the randomness. The inference can extend to forecasting and prediction. This may also include extrapolation and interpolation of time series or spatial data.

Analysis of variance (Anova), chi-square test, correlation, factor analysis, Mann-Whitney U, mean square weighted deviation (MSWD), regression analysis, student t-test, time series analysis, k-nearest neighbors (*k-nn*), majority classifier, group method of data processing algorithms, logistic regression, uplift modeling, naive Bayes classifier, support vector machines etc. are some statistical tests and procedures used in predictive analysis.

Application of statistical tests and procedures will cause time loss if the number of data is excessive. The use of statistical analysis package programs helps to avoid time loss. A number of statistical analysis package programs have emerged to meet the needs of researchers, students and businesses. In this study, the features of selected licensed and unlicensed programs are presented to the researchers and statistical analysis programs are compared in accordance with this information

2. Software

2.1. Open Source (Free) Statistical Analysis Package Programs

2.1.1. SAS University edition

SAS Analytics has launched a free program called SAS University Edition for higher education in the context of educational support. SAS University Edition is a free edition that includes selected SAS products for learning and teaching statistics and quantitative methods. SAS University Edition provides easy access to statistical software for courses such as research, engineering, social sciences, economics, computer science, trade, medicine and health. SAS University Edition is available in two versions. These are SAS Download that runs locally on your machine and Aws Marketplace that requires internet connection. Both versions work in Windows, Linux and Mac OS X operating systems. The SAS University Edition package also includes access to the SAS Analytics community, where you can access e-learning classes, training videos, and more SAS resources. SAS University Edition consists of Base SAS, SAS / STAT, SAS / IML, SAS Studio and SAS / Access components.

Base SAS makes programming with SAS programming language, output delivery service graphics (ODSgraphics) and reporting procedures. It also provides convenient data analysis including intuitive 4GL and easy-to-learn syntax, including correlations, intuitive statistics, distribution and table analysis. SAS Macro Facility reduces coding to modularize work such as easy reuse and maintenance and support for Structured Query Language (SQL).

SAS / STAT offers analysis such as variance, Bayes, categorical data, clustering, discriminant distribution, group order design, multivariate, psychometric, regression, spatial. It also offers methods and techniques such as descriptive statistics, definite methods, group order design, market research, missing value estimation, power and sample size.

SAS / IML provides a matrix programming language for more specific analysis and data research such as matrix functions, linear algebra and statistical functions, time series functions, interactive data analysis, etc..

SAS Studio provides hundreds of SAS statements and procedures along with programming tools that are useful for improving productivity by using autocomplete for built-in syntax help, using the browser-based interface for basic data processing and basic statistical tasks to build SAS code.

SAS / Access provides easy connectivity with SAS and other file formats for access and integration (Anonymous, 2018a).

An intuitive interface, a powerful programming language, comprehensive and reliable tools, a robust and flexible matrix programming language and a great user community are among the benefits and best features offered to users by SAS University Edition.

2.1.2. Paleontological statistics (PAST)

Paleontological Statistics (PAST) is a free statistical software for data analysis, graph creation, data manipulation, univariate and multivariate statistics, time series, ecological analysis, morphometric, stratigraphy and spatial analysis. PAST provides an easy-to-use analysis package that includes statistical, graphical and modeling functions to users. PAST is easy to use because it provides an interactive user interface, easy coding, and a spreadsheet type data entry form where excel users can enter their data. Users can make curve fitting in many types. Linear with bootstrapping and permutation, Breusch-Pagan and Durbin-Watson tests and generalized linear model that includes Logit regression, log-log, logistic, polynomial, lin- log, Von Bertalanffy and sum of sines are some of the curve fitting types located in PAST. PAST provides a good platform for users to perform multivariate statistics including principal coordinates, key components, non-metric and multidimensional scaling, and much more. PAST also offers time series analysis features including spectral analysis, cross-correlation, autocorrelation, Walsh transform, wavelet transform, runs test, short time Fourier transform and Markov chains and Mantel correlogram and periodogram. Directional statistics, angular mean with confidence interval, kernel density estimation of point density, rose plots, alignment detection and more are the geometric analysis methods offered by PAST to users. Analysis of parsimony that includes Heuristic algorithms, Fitch and Dollo characters and Wagner is also performed by PAST. PAST also has a biostratigraphic property and guiding feature including thin plate spline and moving

average (Anonymous, 2018b).

Curve fitting, multivariate statistics, time series analysis and geometrical analysis are the best of the features offered by PAST to users.

2.1.3. MacAnova

MacAnova was written and developed by two statisticians from the Statistical School of the University of Minnesota, Gary W. OEHLERT and Christopher BINGHAM. MacAnova, unlike its name, is not designed for Macintosh computers only and does not perform variance analysis only. MacAnova has several abilities, including analysis of variance and related models, matrix algebra, time series analysis and exploratory statistics. It is reasonably fast, it can be expanded via macros, the data in the spreadsheet can be imported directly and the results can be transferred to other programs via clipboard. MacAnova has a functional or command-oriented user interface. It also offers other features via menu / dialog / mouse type interface. MacAnova is similar to S and S-Plus statistical programs, and this similarity makes it much easier to translate S-plus and R codes into MacAnova. MacAnova codes have various differences from S-Plus and R codes and work well with named data sets and macros libraries. Although MacAnova is not a comprehensive package, it is flexible and powerful enough to be used in research-level statistical calculations, and can also be used extensively in course work. Students can solve statistical and algebraic problems with a few commands. Information and documents about MacAnova, including various versions of the program, are available on the University website. MacAnova provides online usage summaries and full help for more than 550 topics in 8 files (Anonymous, 2018c).

Variables and operations, descriptive statistics, linear and generalized linear models, matrix algebra, random numbers and probabilities, graphs, transformations, in and out data transfer, programmability, dynamic user functions, online documentation, experiment design, time series and multivariate analysis are among the features offered by MacAnova to the users.

2.1.4. GNU PSPP

GNU PSPP is a program for statistical analysis of sampled data and is an alternative to IBM SPSS Statistics and written in C language. PSPP has a graphical user interface. T tests, descriptive statistics, Anova, linear and logistic regression, clustering analysis, reliability and factor analysis, non-parametric tests and more are among the analyzes, methods and techniques offered by the PSPP. Features of the PSPP include more than 1 billion files and variable support, syntax and data files compatible with SPSS, terminal or graphical user interface selection, text selection, postscript, pdf, open document or html output formats and database resources. It has the ability to open, analyze and arrange two or more datasets simultaneously. PSPP supports operations such as loading and saving data and syntax

files as well as having popup windows that provide access to supported statistical analysis and transforms. Features can be used through interactive dialog boxes that display each command's options and required parameters. PSPP can generate high-quality graphics (box and whisker plots, normal probability plots and histograms) to help visualize the distribution of data. These graphs complete descriptive statistics and help determine the most appropriate analysis for the dataset and / or which transformations are required. The PSPP's data selection capabilities (from a subset of variables or from which data only match certain criteria) make it easy to create graphics. Recoding and manipulation of the data using the transformations of the PSPP can be accomplished quickly. These transformations provide simple boolean criteria, arithmetic expressions and mathematical functions without requiring an extra iteration. PSPP also supports many math functions such as random number distrubitions, trigonometry and date-time conversions (Anonymous, 2018d).

Free license, does not expire, enter an unlimited number of samples, import data from any database, two or more datasets run at once, speedy data analysis and core package contains complete functions are some of the benefits provided by the GNU PSPP to the users.

2.1.5. Develve

Develve is a statistical software that helps in quick and easy interpretation and analysis of scientific and experimental data in a technical environment and prevents false assumptions. In short, facilitating and accelerating statistics, and suitable for inexperienced users but sufficiently developed in demanding users. Develve does not have a secret menu. Everything can be accessed directly to improve productivity and results can be seen directly. For example, result graphs can be shifted up and down, and a larger version is opened when the graph is clicked. Develve clearly states that the sample size is large enough to avoid false assumptions when the two variables are distinctly different. Develve helps to create a test matrix for experimental design, to detect if there is a factor that is not balanced, and to develop a robust product at high quality. This makes Develve the perfect six sigma toolbox. Six sigma methodology is a discipline and data-driven approach to maximizing the quality level and minimizing defects in the manufacturing process. The six sigma tools supported by Develve are process capability statistics (cp/cpk) % out of tolerance, regression analysis, correlation, one way Anova, Gauge R&R measurement system analyses, chi square test respectively. Develve also supports a variety of graphs such as histogram, control chart, scatter correlation plot, time series plot and individual dot plot.

Difference, t-test, wilcoxon-mann-whitney test, variation F test, variation Levene test, Anderson Darling normality test, correlation test and regression are among the features provided by Develve (Anonymous, 2018e).

Tests normally distributed data, Tests non normal distribution data, measures system analyses, reliability and offers a statistical handbook are some some of the benefits provided by Develve to the user.

The free version is available for non-commercial purposes and is the same as the paid version. It is foreseen that non-commercial use will primarily involve teachers and students who may wish to use it without any commercial advantage or financial gain.

2.2. Licensed Statistical Analysis Package Programs:

2.2.1. IBM SPSS modeler

IBM SPSS Modeler provides a range of analytic solutions to daily business problems such as text analytics, asset analysis, social network analysis, automated modeling, decision management and optimization. IBM SPSS Modeler has the ability to conduct analysis wherever data is stored, regardless of format and structured status. IBM SPSS Modeler is available in three editions.

IBM SPSS Modeler Gold: It enables organizations to build forecasting models and place these models directly into business processes. This has been achieved through decision management.

IBM SPSS Modeler Premium: It offers a number of advanced algorithms besides capabilities such as text analysis, asset analysis, social network analysis, and automatic modeling and preparation techniques.

IBM SPSS Modeler Professional: It provides an advanced set of algorithms, data manipulation, automatic modeling and preparation techniques to users for creating predictive models and revealing hidden patterns in structured data.

Interactive easy to use interface without the need for programming, automated modeling and data preparation capabilities, access to all structured and unstructured data from different sources, natural language processing (NLP) to extract concepts and sentiments in text, entity analytics ensures the quality of the data and results in more accurate models and integration to Cognos, Netezza, Infosphereand System Z are some of the features offered by IBM SPSS Modeler to the user. IBM SPSS Modeler increases performance by reducing data movement with client server architecture. Integration with other open source technologies such as R, Python, Spark can be achieved to increase the analytical power of IBM SPSS Modeler. Auto classifiers, decision trees, logistic, SVM, time series, anomaly detection, APRIORI, carma, sequence, auto clustering, k-means etc. algorithms are included in IBM SPSS Modeler (Anonymous, 2018f).

Benefits provided by IBM SPSS Modeler to the user include range in advanced algorithms, improved decision making, geospatial and text analytics, automated modeling, gained predictive accuracy and variety of working premises.

2.2.2. Matlab

Matlab is the easiest and most productive software for engineers and scientists. Matlab combines a high-level language with a desktop environment for iterative engineering and scientific workflows. Matrix-based Matlab language is the most natural way of expressing computational mathematics. Linear algebra in Matlab also looks like same linear algebra in the textbook. This directly catches the mathematics behind the user's thinking, which means that the code is easier to write, read, understand and maintain. Matlab is known for its excellent numbers in the numerical analysis research community. Mathematical operations are distributed to multiple cores on the user's computer and all codes are compiled just-in-time with optimized library calls. Matlab allows writing algorithms in parallel by changing for-loops into parallel for-loops or by changing standard arrays into GPU or distributed arrays. These parallel algorithms can be run infinitely scalable in public and private clouds without the need for code changes. The Matlab language also provides users with the integration of flow control, error handling, object oriented programming, unit testing and source control, which are features of traditional programming languages. Integrated tools allow less time to evaluate more ideas. Matlab has 2D and 3D drawing functions for visualizing the results. Matlab supports languages C / C ++, Java, .NET, and Python and languages for embedded systems. In addition to being able to place Matlab code easily in Hadoop systems, Matlab is an important part of model based design, which is used for simulation of physical and discrete events, multi domain simulation, and verification and code generation (Anonymous, 2018g).

The best features Matlab offers to its users include high-level language for scientific and engineering computing, desktop environment tuned for iterative exploration, design, and problem-solving, graphics for visualizing data and tools for creating custom plots, C ++, java, .NET, Python, SQL, C ++, and many other domain-specific tasks, add-on toolboxes for a wide range of engineering and scientific applications, Hadoop, and microsoft excel.

2.2.3. Minitab

Minitab is one of the world's leading developers of statistical software and software for six sigma quality improvement projects. Thousands of companies and over 4000 colleges and universities all over the world use Minitab. Toshiba, DuPont, Boeing, Royal Bank of Scotland, Nestlé and Pfizer are just a few of Minitab's customers. Minitab's products are supported by a variety of services, including training and free technical support. Two new analyzes include experiment design (DOE) and multiple regression in Minitab's assistant, which offers many functionality and enhancements to facilitate statistical analysis and provide more information on quality improvement processes. The user interface allows the user to quickly determine the predicted variables and also the automatic model selection makes

it easy to identify the important variables. With its extensive graphical options, it provides more ways to visually explore the results, while response optimizer makes it easier to find the optimal settings for process variables. Improved experiment design capabilities are enable users to more effectively identify the factors and interactions that affect their processes. Response optimizer can be applied to general factorial designs, and response surface designs can include categorical factors and also automatic model selection can be used for both factorial and response surface designs. Bubble plot, poisson regression, outlier tests, tolerance intervals, stability studies, equivalence tests are some of the other features offered by Minitab (Anonymous, 2018h).

Features of smart data import, automatic graph updating, seamless data manipulation, effortless presentations, basic statistics, regression and Anova, quality tools and design of experiments are among the best features offered by Minitab to users.

Monitor processes over time and evaluate their stability, determine a product's lifetime characteristics, assess how well processes meet specifications, and determine if measurement systems are adequate are some of the benefits provided by Minitab to the users.

2.2.4. Stata

Stata is a complete, integrated statistical software that provides everything for data analysis, data management and graphics. Stata's data management features allow users to combine and reshape data sets, manage variables, and collect statistics across groups or copies. Users can work with up to 2 billion characters with byte, integer, long, float, double, and string variables. Stata also has advanced tools for managing specific data such as survival / duration data, categorical data, survey data, panel / longitudinal data, multiple-imputation data and time series data. Stata makes it easy to create graphs with unique publication quality. Users can generate hundreds of graphics in a repeatable way by writing code, and can output PNG for publishing on the web and output in PDF format for viewing. With the graphical editor integrated into Stata, users can modify graphics. Mata is a complete programming language that compiles, optimizes and executes what you write into byte code.

Although it is not necessary for users to program when using Stata, it is comforting for users to know that a complete matrix programming language is provided by Stata. Mata provides a complete development environment for producing compiled and optimized codes at the same time while providing an interactive environment for matrix manipulation. Mata offers custom features, processing panel data, performs operations on real or complex matrices and provides complete support for object-oriented programming, and is fully integrated into every aspect of Stata. Stata works on Windows, Mac OS X and Linux / Unix computers. The license is not platform specific, meaning that the user can install the Stata license on any of the supported

platforms (Anonymous, 2018i).

Among the best features offered by Stata are arima, Anova and Manova, linear regression, time-series smoothers, generalized linear models, cluster analysis, contrasts and comparisons and power analysis.

Everything in one package per module, cross platform compatible, matrix programming with Mata, broad suite of statistical features, fast, accurate and easy to use and variety of resources are some of the benefits provided by Stata to users.

2.2.5. AcaStat

AcaStat is a statistical data analysis software that generates tables, cross tabulations, descriptive statistics, correlations, ordinary least squares (OLS) and logistic regression, t-tests, nonparametric tests and more. With Acastat, you can use controls, format variable and value tags; set missing values and validate variables. AcaStat offers different solutions designed for Mac OS X, Windows and IOS. It is also designed to allow fast and uncomplicated statistical analysis. AcaStat allows you to extract, copy, paste and drag-and-drop data from spreadsheets. AcaStat analyzes the data, applies t-test, non-parametric test, regression, correlation, and descriptive statistics and provides a summary for users to analyze the results. Summaries can be compared and analyzed with StatCalc. AcaStat can generate confidence intervals, regressions, descriptive statistics, correlations, price elasticity of demand and queuing theories, and decision tables for further analysis to make decisions. For example, you can add or remove variables according to the evolution of the model until a regression model yields the most appropriate result. A horizontal bar chart for frequencies, a descriptive histogram, a distribution gram for correlation, a residual histogram for OLS regression, and a ROC curve for diagnostic accuracy are five graphical types that can be automatically drawn. The analysis output and its contents can be easily copied for the reports and presentations to be created. Analysis options and settings, selecting and removing variables, selecting the most appropriate test, and saving the output are made from the main panel. Examples of how to use different features and an applied statistical book are available. Applied statistical book and other teaching materials can be downloaded free of charge from the AcaStat website. AcaStat makes data analysis and summary creation a fast, simple and straight forward job. There is also Clarity, a low-cost version of AcaStat for students (Anonymous, 2018j).

Benefits provided by AcaStat to the user include simple but complete software, low-cost, drag and drop variable selection, and charts module.

3. Discussion

IBM SPSS modeler does not include twenty-nine statistical analysis methods such as Lasso regression, copula models, extreme value theory, variance

stabilization, Bayesian statistics etc. In addition to these IBM SPSS Modeler performs ten statistical analysis method in a limited manner including ridge regression, nonlinear regression, naïve Bayes classifier, longitudinal data, and univariate time series etc.

Stata does not include a twenty-one statistical analysis method such as nearest neighbor algorithm, naïve bayes classifier, copula models, markov chains, spatial statistics etc. In addition, Stata performs six statistical analysis methods in a limited manner including ridge regression, lasso regression, stochastic volatility models (continuous case), deterministic optimization etc.

Analysis methods such as ridge regression, naïve bayes classifier, copula models, EM algorithm, and propensity score matching, wavelet analysis; neural networks etc are not included by Minitab or are performed in a limited manner by Minitab.

Acastat includes basic statistical analysis methods such as t test, F test, Anova, point-biserial correlation, logistic regression, OLS regression etc.

Markov chain Monte Carlo methods, expectation-maximization (EM) algorithms, diffusions, Markov chains etc. statistical analysis methods are performed by Matlabwith more than ten lines of code writing and also, all other statistical analysis methods can be performed with Matlab.

MacAnova does not include many of the advanced statistical analysis methods including bootstrap and jackknife estimations, survival analysis, quality control, reliability theory, diffusions, meta analysis, roc curves etc.

Develve does not include many of the advanced statistical analysis methods including bootstrap and jackknife estimations, survival analysis, quality control, reliability theory, diffusions, Meta analysis, roc curves etc. Although it is same as paid version, Develve includes basic statistics, design of experiments, measurement system analysis and some other features.

GNU PSPP does not include many of the advanced statistical analysis methods including random forest classification, support vector machines (SVM), deterministic-stochastic optimization, Meta analysis, signal processing, lasso regression, markov chains, time series etc. In addition, GNU PSPP is known as the open source version of IBM SPSS Modeler and has fewer features than IBM SPSS Modeler.

PAST has a wide range of statistical analysis methods, but it does not include some of the advanced statistical analysis methods including

Bayesian statistics, variance stabilization, roc curves, extreme value theory, robust estimation, propensity score matching etc.

SAS University Edition does not include twenty-nine statistical analysis methods such as naïve Bayes classifier, extreme value theory, variance stabilization, diffusions, markov chains and hidden markov models. In addition to this SAS University Edition performs 12

statistical analysis methods in a limited manner including ridge regression, nonlinear regression, naïve bayes classifier, longitudinal data, and univariate time series etc.

With these informations, Matlab, Minitab, Stata, Acastat and IBM SPSS Modeler compared in terms of number of features and functionality offered to users, Matlab performs better than the other statistical analysis software.

With this information, MacAnova, Develve, GNU PSPP, PAST and SAS University Edition compared in terms of number of features and functionality offered to users, SAS University Edition performs better than the other statistical analysis software. This is a predictable result, because SAS University Edition is a version of paid software. Moreover, it has the same features except for a few features.

4. Conclusions

In this study, when looking at the features and facilities provided by the statistical software examined in both groups Matlab may be preferable more than others by the users. If Matlab compared to the other software, it includes almost all of the statistical analysis methods and also the convenience and benefits offered by the paid software. Users can choose the most appropriate paid or free software from the analyzed statistical software according to their needs.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgements

This research was presented as an oral presentation at the International Congress on Domestic Animal Breeding Genetics and Husbandry (ICABGEH-2018) held on 26-28 September 2018 in Antalya.

References

Anonymous, 2018a. SAS University Edition. https://www.sas.com/en_us/software/university-edition.html (access date: 20.07.2018).
Anonymous, 2018b. PAST. http://folk.uio.no/ohammer/past/, (Date of access: 15.08.2018).
Anonymous, 2018c. MacAnova. http://www.stat.umn.edu/macanova/, (access date: 09.08.2018).
Anonymous, 2018d. GNU PSPP. https://www.gnu.org/software/pspp/, (access date: 01.08.2018).
Anonymous, 2018e. Develve. https://develve.net, (access date: 10.07.2018)
Anonymous, 2018f. IBM SPSS Modeler. https://www.ibm.com/analytics/spss-statistics-software, (access date: 26.08.2018).
Anonymous, 2018g. Matlab.

https://www.mathworks.com/products/matlab.html, (access date: 20.08.2018).

Anonymous, 2018h. Minitab. http://www.minitab.com/en-us/, (access date: 05.08.2018).

Anonymous, 2018i. Stata. https://www.stata.com, (access date: 25.07.2018).

Anonymous, 2018j. Acastat. https://www.acastat.com, (access date: 25.08.2018).