J. Statistics
1996(1), 15-28

SOFTWARES FOR STATISTICS

By

AHMED FAISAL SIDDIQUI

(Department of Statistics, Islamia University, Bahawalpur.)

ABSTRACT

A comparative study of various software packages for statistics available in software markets, is attempted. Detailed discussion is made on each selected statistical package with reference to its advantages, merits, software & hardware requirements and demerits with the basic idea to enable a statistician to choose the package best for his range of problems.

KEY WORDS

Personal Computer, Software Package, WINDOWS, DOS, SPSS, Statistica, Statgraphics, Unistat, S-Plus

1. INTRODUCTION

Today, computer has become an indispensable part of our live. In every sphere and department of our businesses, it has made its presence permanent and inevitable. For statisticians, its position is even more important, as the statistics needs both of its fundamental operations of memory & retrievability much more desperately. Interesting enough, it was a statistician, Dr. Herman Hollerith (1877), who invented the punched card computer system, the precursor of today's modern electronic computer, to improve the compilation and processing of the US population census data.

Even today, computer is here with much more helping tools to assist statisticians in their data compilation, presentation, and inference tasks. Software markets are replete with numerous types of statistical packages. Different packages focuses at different sets of problems or take the same problems with a different angle more appropriate to them.

This paper is written with the sole intention to introduce some worthy statistical packages, at the first hand, and to compare them with respect to their functions and capabilities, at the second, so that the existing confusion at the time of their purchase can be curtailed.

It should be born in mind that the paper is not a helping guide, or tutorial, for these packages, it only introduces and then compares them. Another important thing to note that I have selected the latest versions of these packages, rather than to discuss them in general. These latest versions, as the contemporary software, require WINDOWS environment. They will not function properly in the old days' DOS environment. Secondly, most of the software is available, now a days, on CD ROMS, but fortunate enough their disk versions are also available.

2. SPSS (VERSION 6.1)

For as long as anyone can remember, Statistical Package for Social Sciences (SPSS) has been one of the top statistical software package available for professional usage. With an accent on survey analysis, the SPSS slogan "Real Stats, real easy" does indeed describe the philosophy of this program in which many of the most advanced and up-to-date statistical tools and techniques are available as click-on options of the mouse. Furthermore, the package seems to facilitate more on the database market research, analysis of customer mailing list response rates and other aspects of consumer behavior.

Fortunately, or unfortunately for some users, this new version of SPSS works only in WINDOWS environment. The basic SPSS system consists of a shell with everything necessary from the data input, to the preparation and manipulation of data, and even to the coding, inferring and analysis of the data. Analysis of multivariate date is also there upon the same shell. The graphics feature includes more than fifty types of charts (if I am not wrong).

SPSS version 6.1 provides full 32 -bit data processing system and makes extensive use of available toolbars. There are large, clear, easy-to-use dialogue boxes and the mouse's click-on facility to recall any of the pervious one. Data files can be imported, and exported as well, in a wide

range of formats including that of EXCEL (.xls), Lotus (.wk), dBase (.dbf), the standard ASCII field (.txt) and the tab-delimited. This is an added merit of this package. Together with the SPSS developers kit module, it offers full OLE2(on Line Exchange), DDE (Dynamic Data Exchange), WINDOWS' API, ODBC and SQL connectivity. Other modules are applied to this WINDOWS shell as integrated building blocks of a single entity.

The range of modules available (available separately) with SPSS covers many aspects of advanced data analysis and document presentation. The *Tables module,* for example, is almost completely self explanatory and allows the preparation of automatically formatted, tabulated inferred summary or a frequency distribution ready to be exported to any word processor or presentation package. The Experimental Design module covers full range of designs and designing techniques starting from one way analysis of variance to the quite intricate neighbor designs. The *professional Statistics* module provides multivariate data analysis techniques like the Discriminant analysis, the Factor analysis, the Principal Component analysis, the Cluster analysis, and the procedures like Logistic Regression. The Advanced Statistics module includes special routines for medical and social science data analysis such as Survival Analysis, the Multivariate analysis of variance (MANOVA), etc. The Survival Analysis methods include implementations of life tables, the Kaplan-Meier, and the Cox's regression with the option of time dependent covariance analysis.

Another important and very useful module is CHAID (chi-square automatic interaction detector) module. It is used to identify database sub groupings as an alternative and is particularly suitable for use with categorized non-parametric data. Using CHAID, upto *64,000* variables can be analyzed, with result presented in a GAINS chart format - another advance feature of SPSS 6.1, which enables data results to be interpreted and then implemented at a strategic market level. The *Mapinofo* module, another useful module, presents a way of displaying direct marketing, or survey analysis data which includes subject's post codes on a map. Data can, either, be displayed symbolically, or by color codes on map that can display boundaries, urban areas and even the street names.

The user needs a certain degree of expertise to select his requisite module, as the user has to purchase them separately. The only disadvantage with this package is its price.

In short, SPSS 6.1 is the top most statistical package available, if one can afford it.

3. STATISTICA (VERSION 4.5)

This is a package seems to be designed specially for professional statisticians. All of its modules are floating in nature and you can have six or more open at the same time, each with a different spreadsheet, each with a separate data set, and each with some other aspect of the analysis. Additionally, all these can be layered, tiled or iconized, just as the Window's main screen.

The comprehensive range of available statistical analytical functions spans from the correlation analysis to the nonparametric tests, the multiple and non-linear regression analysis, from canonical analysis to the process analysis, from experimental design to multidimensional scaling, cluster analysis, which are multivariate analytical techniques, and much much more. Available with nineteen probability distribution functions and an instant probability calculator, this package is no slouch and the inherent capability for multitasking, no doubt, facilitates the background printing of graphs, charts and tables.

The radiant feature of STATISTICA is its unmatched range of graphs and charts, as it is claimed, to feature over 600 graph and charts. These include some very special and rarely seen features such as hanging histobars, which can be used as a visual test for normality, and the categorized normal probability plots which examine normality aspects of group homogeneity. Moreover, in graphical aspects of the curve fitting analysis, it is quite possible to plot a non-linear curve on to box-plots of the time series data. Categorized data can be represented as vertical slices (panes) which display the relative data spreads on a spectral graph. I have not seen such novel features in any other statistical package as yet.

Another feature of STATISTICA is the representation of more than one probability distribution through multiple histograms on a single graph sheet. All these graphs are completely customizable. Edited formats can be saved as templates, and graphic can be saved as templates, and graphic can be saved in a choice of formats or used as embedded objects - for example, STATISTICA provides OLE2 (on line exchange) linkage to MICROSOFT WORD 6. Most of the statistical information obtained can be arrived at by purely numerical analytical methods but the advanced graphical routines available in this package enable a user to actually see what is happening. There are times when a visual portrayal of information have worth more than a collection of numerical results.

STATISTICA is widely used in US Government statistics departments (PC word, August 95) and one can well imagine a situation when a user of this package is informing 12 different callers that he has their data on screen right now. Additionally, this feature enable on screen comparisons to be made between different data sets running the same procedures in different WINDOWS simultaneously. The probability calculator, an option unique to STATISTICA, is especially useful for such comparisons - the whole package can be personalized to suit the style a user wants. Menus, floating menus and toolbars can be edited and reshaped, hot - keys can be defined, and redefined, or used in conjunction with macros. These macros may also be used to enable other packages, such as Microsoft's WORD or Lotus's AMI PRO, to call specific module routines, using the STATISTICA's own command language. The chief advantage of its independently functioning module construction is that a module can be called without its base, thereby reducing memory overhead and permitting DDE (Dynamic Data Exchange) access to other programs, for example EXCEL form AMI PRO, without exceeding an *8* MB RAM limit.

STATISTICA is well documented with four volumes of handbooks totaling more than *13,000* pages. There exists also on-line help, in addition to usual help files, which is no less than a statistical advisor to suggest what to do and which methods is best to use in a particular problem. In addition to the full STATISTICA package, Stat-Soft also distributes Quick STATISTICA (for WINDOWS) which includes the

basic Statistical analytical functions with full data management and graphics capabilities.

In short, it is a full featured statistical computer package available in the markets. I think its most frightening features is its usage which demands not only proficiency but experience too. The package can not be labeled as easiest to use. It is better to use its free demonstration kit before purchasing it.

4. STATGRAPHICS (VERSION 1.1)

Among those statistical computer packages available for WINDOWS environment, STATGRAPHICS is the most easy to use. Basically, it is a statistical package having immaculate graphic routines and specially designed charts to prepare graphical presentations of compiled information. There are clearly and self-explanatory dialogue boxes, simple toolbar menus and fast graphics. A little play with mouse will result in perfectly formatted output of a almost every available statistical analytical routine. For example, when working with a column of data, the drag and drop click-on graphics feature enables descriptive histograms, scatter plots, symmetry plots and such like to be produced in an instant. Another mouse click will select any diagrams which is to be examined separately, discarded, edited or saved and printed.

Having selected *Distribution Fitting,* for example, clicking the right button of your mouse will provide a list of five probability distribution functions (Normal, Log Normal, Exponential, Extreme value and Weibull) to chose form. Once selected the probability distribution function, a click-on tabular options, with selection of all, gives Chi-square and Shapiro Wilkes tests for normality, goodness of fit, tail areas and critical values, automatically formatted and ready to print for the previously selected probability distribution function, as well.

This feature certainly makes the package easy-to-use, and of with high quality output, as it is claimed by its designers, but "Are only five probability distributions sufficient?" Secondly, the package does not provide, an almost omnipresent, feature to sketch them all at the same time on a single histogram of the data. Similarly, there is much to be

said about the goodness of fit tests provided in the package but the comparative evaluation approach incites for the lack of exhaustively objections. Ease of use and simplicity has its price and in this case it is the expense of available functions and flexibility.

This some click-on and easy approach applies to other statistical analytical functions of the base module, such as analysis of variance technique (ANOVA), regression & correlation analysis, categorical analysis and choices for data plotting routines - all give the user the always cherished option to click-on and identify individual points on graph which makes it easy to identify *outliners* and important clusters portraying individuals personalities in the data sets.

The analysis of variance (ANOVA) routine caries two options, i.e., the single factor designs and the multi factor designs. The multiple regression analysis produces correlation matrices, procedure summaries; including conditional sums of squares, and ANOVA tables for the regression analysis, in addition to indicating the unusual residuals and influential points. Here, the click-on graphics option includes component effects, observed or residuals versus predicted, row number and X, and the interval plots.

The most promising modules of the package are the *Quality control* module and the *Time Series* module. There is no compromise, at all, on the number of features available in these modules. The Quality control module has almost every conceivable quality control method with *16* types of analysis ranging from X-bar and the R-charts through process control, OC charts, to tolerance analysis and much more.

Both of these modules are exquisitely documented with their own handbooks designed as tutorial reference guides. Someone with only a passing knowledge of statistical forecasting or quality control can use these references to build the basis of a working familiarity and knowledge of the techniques described. At the top, listing of available features takes several pages and only an experienced quality control or forecasting specialist would be able to evaluate such a combination of analytical tools. The computer interface, like the base module, is quite excellent and the integration is flawless.

In terms of click-on ability, this package is one of the best WINDOWS applications available in any category. But the limited range of functions is a disadvantage unless you had the good fortune to be quality consultant or were involved in prediction and forecasting.

In short, the STATGRAPHICS is fast to learn and fairly easy-to-use. Ignoring its limiting range of probability distributions function, the package is exquisite.

5. UNISTAT (VERSION 4.0)

UNISTAT, an age old renowned statistical package, has many new features in its latest version designed specially for the WINDOWS. The new statistical analytical routines a available in version 4.0 include enhanced experimental design option, comparison of regression slopes and intercepts, nonparamentric multiple comparison tests for Friedman two way analysis of variance, the Quade tests, and multiple comparison tests for medians and variances. It provides, also, a feature of interactive selection terms for unlimited number of factors in analysis of variance technique. The list of new features include, also, the option of spreadsheet selection of data columns for analysis and exquisitely redesigned dialogue boxes in the style of Microsoft's Wizards offering intelligent prompts. For the users of other spreadsheets and editors, the data processor retains its menus and allows data to be posted from lotus 123 or any other WINDOWS spreadsheet, and through DDE (dynamic data exchange) permits a two way interchange of data with other programs. An on-line database connectivity module is also included in the package to provide compatibility with other software's.

All statistical analytical routines can be accessed either via a choice of WINDOWS pull down menus or from the UNISTAT main menu, as the traditional UNISTAT did, remains as a click-on alternative at the far right of the toolbar. The main menu band in this latest version includes *undo* and *repeat* like functions, and the choice of output control is a simple Microsoft - style item adjacent to the main menu on the toolbar.

UNISTAT 4.0 provides OLE2 access to Microsoft's EXCEL as a means of achieving formatted and styled output. The prime advantage is that

the user, no longer, has to send output to a fixed width font text editor, as in an old mainframe version of the UNISTAT used to. In the past, the output printout were often included in a folder at the end of a main presentation. And if you need a summary of the data, with graphs, charts and spreadsheets, right in the main body of the document, you would require pasting and formatting, even with a packages offering OLE2 import facilities you would require the chart, graph, or spreadsheet to be named and saved, as well, in advance, which may be more time consuming than the actual data analysis.

With UNISTAT 4.0, however, selected output can be automatically formatted into Microsoft's WORD or EXCEL Type tables, without going through all old fashioned fatigues. Graphics, charts and spreadsheets can be exported to MS editor tools can be used for the final formatting. Secondly all these graphics editing, and formatting as well, can be carried out form within the drag-and drop UNISTAT graphics editing environment. you can call UNISTAT from EXCEL to perform advanced analysis upon a data object, with on output formatted directly back into EXCEL as the host application. Doesn't it call beauty?

High quality documentation for on the spot help is available with full details of all the available statistical algorithms in the package. UNISTAT 4.0, as its previous versions, can be used as a compact standalone program in WINDOWS 3.1 on a *386* system with 4MB of RAM. OLE2 implementation outputting to either Microsoft's WORD or EXCEL runs effortlessly on a *486* system with 8MB or RAM. Whereas full OLE2 implementation outputting to EXCEL and WORD simultaneously requires 16MB or RAM.

UNISTAT 4.0 works with 19 probability distributions functions, enabling data histograms to be plotted with fitted probability distribution such as Normal or Student's t, everyone superimposed over the same histogram. In this way, six probability distribution functions can be plotted, and superimposed as well, on the original histograms, enabling a visual comparison of the optimum fit. The range of statistical analytical routine available in UNISTAT includes Multivariate and Cluster analysis, Discrimenant analysis, Principal Components analysis, Factor

analysis and a variety of multivariate plots. The time series analysis uses ARMA (outoregressive moving average process) with Browns, Holts, and Winters routines, while the Survival analysis includes life tables, Kaplan-Meier analysis and Lee Desu comparison like statistical techniques. In the field of Quality control procedures, there are fifteen types of control charts and the package even offers Furrier transformations. All these features are enough to rank UNISTAT as excellent value-for-money package both for beginners and experiences statisticians.

Anyhow, it is a value-for-money package which does not require additional modules, another radiant feature, and will produce excellent outputs. However, among its demerits the example of *"no single click-on graphic of observed y-fitted values"* can be given.

6. S - PLUS (VERSION 3.2)

It is purely a professional statistical package and if you are looking for a simple, quick and easy-to-use WINDOWS statistical package with everything available as a click-on options, then simply forget S-Plus. The package is equipped with almost everything you want from a statistical package and if you are ready to invest your time to learn some novel data analysis techniques, then you might find that S-Plus is the best for you.

S-Plus claims to set new standards in data analysis. It is a rich graphical data analysis system. It takes up over 20MB of your hard disk space and comes with six volumes of documentation totaling over *2,000* pages, and more than *1600* statistical analytical functions. It offers some fantastic functions, for example, *least median regression fit,* which others don't even have heard of, and facilities for tackling complex numbers, Vectors with their peculiar geometry, matrices and other aspects of modern mathematical accretion not normally found in a statistical computer package.

Although the package S-Plus is not easy-to-use, however its real strength lies in the variety of analytical routines available. The range of statistical routines and functions is quite impressive by any standard and is, no

doubt, the most puissant in comparison with any other statistical packages reviewed and tested in this paper. With the only exception of some high level econometric analytical functions, S-Plus have virtually every conceivable statistical routine, or its equivalent, offered by any other statistical package, reviewed here. For instance, it produces a binary response tree, a far superior diagnostic to the GAINS chart of the CHAID (chi-square automatic interaction detector)module available with SPSS 6.1. Similarly, it can also map any indexed data with a brush option to paint in any color. Other features of the package include Cluster analysis, a multivariate analytical technique, and many other modern regression methods. The range of Variance analyses, Time series and Survival analysis is equally impressive.

The package is provided with a developer's tool kit, too, and a *450* pages Programmer's Manual with routines for loading dynamic link libraries into main system. The command language set is a functional language which evaluates each function call in an expression, in a separate frame of memory. The package is supplied with a library of development tools which can interface with *FORTRAN* or *C* to produce object oriented specific applications within the WINDOWS environment. It does not need any additional modules, but there is an interface module to run the package on Novel network with a module to interface S-Plus with the Maple Keme.

Potential users of S-Plus would be those who have an experience of using other statistical packages and are familiar with the object oriented programming languages. Despite this being a WINDOWS' program, the system does not have an exhaustive pull down menus, or toolbars or not much to clock-on from mouse. The reason is that the available ten pull-down menus each offering ten feature, covers less than ten percent of the available *1600* statistical functions and techniques. The existence of multiple menus also give rise to the problems of navigation, hence the strict adherence to the command line interface is required. S-Plus is purely a command line driven package, even more than Mathematica, and is closer to Matlab than to other WINDOWS statistical package reviewed here.

The only demerit of the package is its price. Anyhow, even with a poor interface, and clumsy command line instruction set it is a good statistical package. But beware, this package is not meant for beginners.

7. CONCLUSIONS & RECOMMENDATIONS

I have taken five statistical computer packages, in this paper, for their comparative study. Software users are quite justified to expect the same assurance of overall quality, consummation, and value-for-money as any other. Unfortunate enough, this is not the case.

SPSS 6.1 is no doubt an excellent statistical package with a comprehensive variety of statistical analytical techniques and an easy-to-use organization, it cannot be seen and understood as offering satisfactory value-for-money to deserve the ultimate software accolade. Similarly the speed and the memory requirements make its selection for the number one statistical software available in the markets quite dim. STATGRAPHICS 1.1 also stand at a high echelon for its easement. Unfortunately, this fast, easy to use WINDOWS application has not enough modules available and this restricts its orbit too much to put it in the running for the number one position.

On the basis of value-for-money and statistical analytical features, I was left with a dissension between STATISTICA 4.5 and UNISTAT 4.0. Both of these are well featured programs, providing comprehensive range of statistical analytical and graphics routines, with a choice of functioning demos, the reader is, therefore, advised to try them both before purchasing. As a WINDOWS application, STATISTICA 4.5 is showing its maturity and does not make plenty use of the GUI. UNISTAT 4.0 makes the best use of the WINDOWS environment and is a noteworthy improvement on UNISTAT 3.0, which itself offers excellent value-for-money. Although Arcus 3 and Micro fit 3 provide specialist routines at down to earth prices, they both feature DOS interfaces reminiscent of the late eighties.

So in my opinion, STATISTICA 4.5 and the UNISTAT 4.0 are the statistical packages that can compete for the number one position. It is

quite difficult to discriminate between them for general use. Both of these packages should be with a statistician.

REFERENCES

[1] **RALSTON, A./MEEK, C.L. (1976)** "ENCYCLOPEDIA OF COMPUTER SCIENCE", *VAN NO STRAND REINHOOD COMPANY, LONDON.*

[2] **BITTER, G.G. (1992)** "MACMILLAN ENCYCLOPEDIA OF COMPUTERS", *MACMILLAN PUBLISHING COMPANY, NEW YORK.*

[3] **GROUP TEST** "STATISTICAL SOFTWARE" PC WORLD, JULY 1995.

[4] "YOUR NEXT OPERATING SYSTEM" *PC MAGAZINES, OCTOBER 1995*