STATISTICAL ANALYSIS IN SOFTWARE PACKAGES*

*by H.F. MAGALIT***

INTRODUCTION

Software in computer centers are developed by (1) in house programing, (2) contract programming, and (3) utilizing proprietary software services. The latter develops software packages which are leased or sold to many users and are offered at lower cost since the cost of development is spread to the users. The trend over the past few decades has been to develop software that reduces both the time and the level of sophistication required to program and to operate a computer system. This trend, that makes computer available to a broader, less experienced group of users, will continue in the future.

SOFTWARE FOR STATISTICAL COMPUTING AND ANALYSIS

For statistical computing and analysis, there are many sophisticated but easy to use software packages that are available at present. Among these are (1) the Statistical Package for the Social Sciences (SPSS) and (2) the Statistical Analysis System (SAS).

Both packages are capable of providing

(1) information storage and retrieval
(2) data modification and programming
(3) file handling – tools for sorting, merging, editing, and updating data sets

*Paper presented at the Third National Convention on Statistics, December 13-14, 1982, Philippine International Convention Center.

**Associate Professor of Statistics, Department of Statistics and Statistical Laboratory, College of Arts and Sciences, University of the Philippines at Los Baños.

(4) wide range of statistical analysis
(5) report writing

Both packages which have wide range of applications, are being developed continually and are supported by established organizations for maintenance. Thus they are easy to maintain locally, and allow inclusion of additional program developed by local users as well as those by the corresponding owners.

Both packages are now used in teaching. SPSS is being used to teach statistics for the social sciences. SPSS has good manuals while the manuals of SAS were written probably for statisticians, since it contains very little of statistical computing.

COMPARISON OF CAPABILITY OF SAS AND SPSS

We will take a close look at the programs and program statements (those enclosed in parenthesis) of both packages and classify them into ten categories as follows:

(1) DESCRIPTIVE STATISTICS

SAS	*SPSS*
CORR	PEARSON CORR
	NONPAR CORR
FREQ	FREQUENCIES
MEANS	CROSSTABS
	BREAKDOWN
RANK	–
SUMMARY	AGGREGATE
UNIVARIATE	CONDESCRIPTIVE
PCTL	–
TABULATE	–
–	MULT RESPONSE
–	NPAR TESTS

(2) ANALYSIS OF DESIGNED EXPERIMENTS (mostly linear models)

SAS	*SPSS*
ANOVA	ONEWAY ANOVA (MCA)
TABLES	–
DUNCAN	–
FUNCAT	–
GLM	REGRESSION
NESTED	–
LATTICE	–
NPARIWAY	NPAR TESTS
PROBIT	–
TTEST	T TEST
VARCOMP	–
–	RELIABILITY
HARVEY	
LOGIST	

(3) **REGRESSION AND MODELING** (nonlinear & linear models)

SAS	*SPSS*
STEPWISE	–
REG	REGRESSION (STEPWISE)
GLM	–
RSREG	–
RSQUARE	–
COXREGR	–
RIDGEREG	–
SYSREG	–
SIMLIN	–
SIMNLIN	–
SYSNLIN	–
NLIN	–
MODEL	–
STATESPACE	–

(4) MULTIVARIATE

SAS	*SPSS*
CANDISC	PARTIAL CORR
CANCORR	CANCORR
CLUSTER	–
FASTCLUS	–
NEIGHBOR	–
TREE	–
DISCRIM	DISCRIMINANT (STEPWISE)
STEPDISC	–
PRINCOMP	–
FACTOR	FACTOR (SCORE)
VARCLUS	–
GLM (MANOVA)	REGRESSION
–	ANOVA (MCA)
GUTTMAN	GUTTMAN SCALE
SCORE	–

(5) TIME SERIES AND FORECASTING

SAS	*SPSS*
AUTOREG	–
ARIMA	BOX AND JENKINS
FORECAST	REGRESSION
STATESPACE	–
X11	–
SPECTRA	–

(6) GRAPHICS

SAS	*SPSS*
CHART	CHART
PLOT	–
–	SCATTERGRAM

(7) INTERFACES

SAS	*SPSS*
CONVERT	MCA
BMDP	–
OSIRIS	–
SPSS	–
DATA-TEXT	–

(8) SPECIAL

SAS	*SPSS*
–	REPORT
SURVTEST	SURVIVAL
PLAN	–
MATRIX	–
TRANSPOSE	–
MORTGAGE	–
COMPUTAB	–
ITEM	–
RANK	–

(9) UTILITIES

SAS	*SPSS*
PRINT	–
CONTENTS	(WRITE FILEINFO)
–	(LIST ARCHINFO)
FORMAT	(RECODE)
–	(VALUE LABELS)
SORT	SORT CASES
(MERGE)	MERGE
	(ADD VARIABLES)
(SET)	(ADD CASES)
(UNIFORM)	SAMPLE
(LAG)	LAG
(OPTIONS)	(OPTIONS)
(UPDATE)	

DATASETS
DELETE
EDITOR
COPY
EXPLODE

(10) **OPERATING SYSTEM UTILITIES**

SAS	*SPSS*
(PUT)	(WRITE CASES)
FORMS	
PDS	
PDSCOPY	
PRINTTO	
RELEASE	
SOURCE	
TAPECOPY	
TAPELABEL	

As can be seen from the foregoing tabulation both software packages have rapid phenomenal growth in the last decade. SAS grew substantially in time series analysis, forecasting and modelling while SPSS has now a complete frange of non-parametric tests. Since SPSS is much more known and widely used presently in the Philippines only the SAS procedures in alphabetical order will then be briefly summarized. The options in each procedure are in capitalized letters.

BRIEF SUMMARY OF PROGRAMS (PROCEDURES) IN SAS

ANOVA – performs analysis of variance for balanced data. MANOVA, FREQ, DUNCAN, BON, ABSORB, GABRIEL, REGWF, REGWQ, SCHEFFE, SIDAK, SNK, LSD, TUKEY, AND WALLER.

ARIMA – analysis and forecasts univariate time series data using the auto-regressive integrated moving average model developed by Box and Jenkins. LAG, FORECAST, RESIDUAL, OUTPUT.

AUTOREG – estimates the parameters of a linear model whose error is assumed to be an autoregressive process of a given order. Should only be used for ordered and equally spaced time-series data with no missing values except at the end. PARTIAL, GINV, COVB, CORRB, LAG, OUTPUT.

BMDP – calls any one of the BMDP, Biomedical Computer Programs, to analyze data in an SAS data set and print the results.

CANCORR – performs canonical correlations, useful in investigating the relationship between two groups of variables. OUTPUT.

CHART – produces vertical and horizontal bar charts (histogram), pies and star charts.

CLUSTER – designed to help identify cluster of observations that have similar attributes, performs a hierarchical cluster analysis, using an algorithm outlined by S.C. Johnson. Should not be applied to data with more than 250 obsrevations. (see FASTCLUS). STANDARD.

COMPUTAB – can produce many types of row-by-column reports like balance sheet and income statements. Can put the data in a desired report format

CONTENTS – prints description of the contents of SAS data sets.

CONVERT – converts SPSS, BMDP, DATA-TEXT and OSIRIS system files to SAS data sets (files). OUTPUT.

COPY – copies OS libraries that contain SAS data sets from disk to disk, disk to tape, or tape to disk. OUTPUT.

CORR – computes correlation coefficients between variables, including Pearson product-moment, weighted product-moment, Spearman rank-order and Kendall tau. Also computes univariate descriptive statistics. SSCP, COV, OUTPUT, WEIGHT, FREQ.

COXREGR – performs regressions on one to twenty independent variables on a censored dependent variable using D.R. Cox's life-table regression model. Alternately, it can perform tests of association between individual independent variables and the censored dependent variable using Dr. P. O'Brien's logit rank procedure, a nonparametric alternative.

DATASETS – can list, delete, and rename SAS data sets contained in an SAS data base.

DELETE – deletes an SAS data set from the disk or tape on which it is stored.

DISCRIM – develops a discriminant model that it uses to classify each observation into one of the groups and then summarizes the performance of this discriminant model. The model is determined by a measure of generalized squared distance. POOL, WCOV, WCORK, PCOV, PCORR, PRIOR, PRIOR PROP, TESTDATA, OUTPUT.

DUNCAN – performs Duncan's multiple range test and the Waller-Duncan K-ratio t-test. WEIGHT.

EDITOR – designed for editing SAS data sets.

EXPLODE – produces oversize printing of text by expanding each letter into a matrix of characters.

FACTOR – performs factor analysis. Principal axis factoring, image analysis, alpha factor analysis and iterated principal axis factoring are all available, along with varimax, equamax, quartimax, and oblique rotation techniques. NFACT, MINEIGEN, PORTION, ROTATE, PREROTATE, EIGENVECTORS, SCORE, PLOT, MAXITER, OUTPUT.

FASTCLUS – is designed for disjoint clustering of very large data sets ($100 \leqslant n \leqslant 10{,}000$) and requires only two or three passes over the data set. MAXC, LIST, OUTPUT.

FORECAST – fits univariate time series automatically and produces forecasts. Uses stepwise autoregressive method and exponential smoothing. METHOD, LEAD, INTERVAL, TREND, OUTPUT.

FORMAT – allows you to define your own output formats giving value labels to variable values. FUZZ.

FORMS – can produce mailing labels, envelopes, external tape labels, file cards and any continuous line-printer forms that do not require page headings.

FREQ – produces one-way to n-way frequency and cross tabulation tables. CHISQ, LIST, WEIGHT, OUTPUT.

FUNCAT – models functions of categorical responses as a linear model. It uses generalized least squares to produce minimum chi-square estimates according to the methods proposed by Grizzle, Starmer and Koch. FREQ, PROB, ONEWAY, X, XPX, COV, COUB, CORRB, PREDICT, NOINT, ML, WEIGHT.

GLM – analyzes general linear models. It handles both nominal and continuous independent variables and generates automatically the dummy variables for the nominal variables. It can be used in many other analyses such as: (1) analysis of variance for unbalanced data, (2) analysis of covariance, (3) response surface models. (see RSREG), (4) weighted regression, (5) polynomial regression, (6) multivariate analysis of variance (MANOVA). ABSORB, CONSTRAST, ESTIMATE, FREQ, LSMEANS, E, STDERR, PDIFF, MANOVA, MEANS, DUNCAN, NO INT, SOLUTION, INVERSE, CLM, ALPHA, XPX, TOLERANCE, BON, GABRIEL, REGWF, REGWQ, SCHEFFE, SIDAK, SNK, LSD, TUKEY and WALLER.

GUTTMAN – creates and evaluates a Guttman scale model for a set of items represented by numeric variables. PROCTOR, WEIGHT.

HARVEY – Least Squares and Maximum Likelihood General Purpose Program (LSMLGP), developed by W.R. Harvey.

LATTICE – computes the analysis of variance and simple covariance for data from an experiment with a lattice design.

LOGIST – logistic regression program using the maximum-likehood method.

MATRIX – is a complete programming language in which operations are performed on entire matrices of values. FUNCTIONS, OUTPUT.

MEANS – produces simple univariate descriptive statistics. KURTOSIS, SKEWNESS, MAX, MIN, USS, T, FREQ, OUTPUT.

MODEL – is used to specify the programming statements that form a model to be processed later by other procedures as SYSNLIN or SIMNLIN. OUTPUT.

MORTGAGE – calculates parameters for a mortgage loan with equal periodic payments and fixed interest rate compounded each period.

NEIGHBOR – performs a nearest neighbor discriminant analysis, classifying observations into groups according to either the nearest neighbor rule or the k-nearest neighbor rule. THRESHOLD, LIST, TESTDATA, TEST LIST, TEST CLASS.

NESTED – performs analysis of variance and covariance for data from a experiment with a nested (hierarchical) structure.

NLIN – produces least-squares or weighted least-square estimates of of the parameter of a non-linear model. It uses the steepest-descent, Gauss-Newton Marquardt method or multivariate secant. PLOT, CONVERGENCE, METHOD, PARAMETERS, BOUNDS, WEIGHT, OUTPUT.

NPAR1WAY –performs a one-way analysis of variance on ranks and certain rank scores.

OPTIONS – SAS options (OBS, FIRSTOBS, ERRORABEND, NOLABEL, BLKSIZE, etc.).

PCTL – computes the percentiles for one or more numeric variables and outputs them to an SAS data set.

PDS – a utility to list, delete and rename the members of a partitioned data set.

PDSCOPY – copies partitioned data sets containing load modules from disk to disk or from disk to tape.

PLAN – generates randomized plans for experiments.

PLOT – graphs one variable against another, producing a printer plot.

PRINCOMP – computes principal components of variables in an SAS data set and outputs them to a new data set. Maybe computed from the correlation matrix or the covariance matrix. OUTPUT.

PRINT – prints a listing of the values of some or all the variables in an SAS data set. SUM, JUMBY, ROUND, ID.

PRINTTO – gives you control over the output of SAS procedures. allows SAS output as input data for another procedure in the same job.

PROBIT – calculates maximum likelihood estimates of the intercept, slope and natural (threshold) response rate for biological assay data. A modified Gauss-Newton method is used to compute the estimates.

RANK – computes rank for one or more numeric variables in a data set. NORMAL, TIES, OUTPUT.

REG – fits least-squares estimates to linear regression models. Can use correlations or cross products as input. WEIGHT, FREQ, RESTRICT, MTEST, OUTPUT (General-purpose procedure for regression).

RELEASE – releases unused space at the end of an OS data set (not just SAS data sets).

RIGDEREG - computes ridge and incomplete principal component parameter estimates for linear regression models. Used to analyze inadequate data that exhibits extreme multicollinearity.

RSQUARE – performs all possible regression for one or more dependent variables printing the R^2 value for each model. If k independent variables are specified, RSQUARE evaluates each of the 2^k-1 linear models.

RSREG – fits the parameters of a complete quadratic response surface, then determines critical values to optimize the response with respect to the factors (2 or more) in the model. LACKFIT, WEIGHT, PREDICT, RESIDUAL, ACTUAL, COVAR, OUTPUT.

SCORE – multiplies values from two SAS data sets, one containing coefficients (for example, factor scoring or regression coefficients) and the other containing the original data used to calculate the coefficients. FACTOR, SCORE, TYPE, OUTPUT.

SORT – can either rearrange the observations in an SAS data set or create a new SAS data set containing the rearranged observations. OUTPUT.

SIMLIN – read the coefficients for a set of linear structural difference equations (usually produced by SYS REG), computes the reduced form, generates predicted values and generates forecasts by simulating the model. Limited only to models that are (1) linear with respect to the parameters, (2) linear with respect to the endogenous variables, (3) square – as many equations as endogeneous variables, (4) nonsingular – the structural coefficients on the endogeneous variables form an invertible matrix. LAG, TYPE, OUTPUT.

SIMNLIN – is used to solve simultaneous systems of nonlinear equations repeatedly. It is designed for solving econometric models to generate predicted or simulated values across a time series. Three solution methods are available: (1) Newton's method (2) the Gauss-Seidel method and (3) the Jacobi method. STATIC, CONVERSE, MAXITER, OUTPUT.

SOURCE – prints or unloads the contents of partitioned data sets containing card-image (80-byte) records.

SPECTRA – produces estimates of spectral and cross-spectral densities of a multivariate time series. (Uses a finite Fourier transform to contain periodograms and cross-periodograms. OUTPUT, WEIGHTS, PH, COEF, ADJMEAN, WHITETEST.

STANDARD – standardizes variables in a data set to a given mean and/or standard deviation. OUTPUT.

STATESPACE – can analyze and forecast multivariate series data. This includes transfer function models that have random inputs. OUTPUT.

STEPDISC – performs a stepwise discriminant analysis by forward selection, backward elimination or stepwise selection. SLE, SLS PRZE, PRZS, TOL, MAXSTEP.

STEPWISE – provides methods for stepwise regression; namely
(1) forward selection
(2) backward elimination
(3) stepwise
(4) maximum R^2 improvement
(5) minimum R^2 improvement

SUMMARY – creates an SAS data set containing summary desriptive statistics. FREQ, T, MAX, MIN, PRT, OUTPUT.

SURVTEST – tests for differences between one or more survival curves. Three tests may be performed: (1) Gehan-Wilcoxon test (permutation based on ranks) (2) logrank test (equivalent to Mantel-Haenzel) (3) likelihood ratio test (exponential model).

SYSNLIN – estimates parameters in a simultaneous system of nonlinear equations using (1) nonlinear two-stage least squares (2) nonlinear three-stage least squares (3) nonlinear iterated three-stage least squares (4) nonlinear ordinary least squares or (5) nonlinear "seemingly unrelated" or joint generalized least squares. DW, WEIGHT, -WEIGHT-' CONVERAGE, MAXIT.

SYSREG – estimates coefficients in an interdependent system of linear equations using (1) ordinary least squares, (2) two-stage least squares, (3) limited information maximum likelihood, (4) three-stage least squares, or (5) "seemingly unrelated" regressions. OUTPUT, OUTSSCP, OUTEST, COVOUT, NOINT, XPX, DW, INVERSE.

TABLES – computes and tabulates simple univariate statistics for all applicable variables for a control group and up to nineteen treated groups.

TAPECOPY – copies an entire tape volume (or files from one or several tapes) to one and only one output tape volume.

TAPE LABEL – lists the label information of an IBM – standard-labeled tape volume.

TRANSPOSE – transposes an SAS data set, changing observations into variables and vice versa. OUTPUT.

TREE – prints a tree diagram known as dendogram or phenogram, using a data set created by CLUSTER or VARCLUS.

TTEST – computes a t statistic for testing the hypothesis that the means of two groups of data are equal.

UNIVARIATE – produces simple descriptive statistics for numeric variables. OUTPUT.

VARCLUS – performs either disjoint or hierarchical clustering of variables based on correlation matrix. The clusters are chosen to maximize the variation accounted for by the first principal component of each cluster. The input may be CORR, COV, SSCP, FACTOR. MINCLUSTERS, MAXCLUSTERS, MAXITER, MAXSEARCH, PARTIAL, WEIGHT, FREQ.

VARCOMP – computes estimates of the variance components in a general linear model.

X11 – may be used to seasonally adjust monthly or quarterly time series (an adaptation of the Bureau of Census X-11 Seasonal Adjustment Program).

SUMMARY

Both SAS and SPSS are capable of providing

(1) information storage and retrieval
(2) data modification and programming
(3) file handling
(4) wide range of statistical analysis
(5) report writing
(6) inclusion of additional programs developed by local users and those to be developed by owners.

Today SAS is used around the globe by statisticians, market researches, biologists, auditors, social scientists, business executives, medical researchers, computer performance analysts and many others. For analyzing social science data SPSS is usually more than adequate and available in most computers, even in minicomputers. SAS, on the other hand, runs only on IBM 360/370 or the latest IBM models (and plug-compatible machines such as CDC, Itel, etc.) under OS, OS/VS, VM/CMS, DOS/VSE or TSO and therefore its use is limited to the above mentioned computers. However, with the availability of a large number of utility programs within its system, SAS has grown into an all-purpose data analysis system not just for statistical computing and analysis.

The computing needs of most researchers can be met by SAS. Instead of learning programming languages, several statistical packages, which are often incompatible, and utility programs, you only need to learn SAS.