Editor's Comments

Copies of the next two papers were sent to the authors of EMOS, SPSS, SAS and OSIRIS for their comments. These comments were then sent to the original authors of the papers for their comments. This, we hope, will eliminate some of the potential for misunderstanding of a paper of this type.
A COMPARISON OF BMD, SAS AND SPSS

by

T. Papaioannou, G.P.H. Styan, and L.L. Ward
McGill University and Bell Canada

Three of the most extensively used statistical subroutine packages are BMD (Biomedical Computer Programs), SAS (Statistical Analysis System), and SPSS (Statistical Package for the Social Sciences). These user-oriented packages possess routines which can perform various statistical techniques. Although they are generally similar in both cost and ease of use, they are often quite different in their scope, content, and presentation. The following notes illustrate some of these similarities and differences. Unless otherwise indicated the results reported below refer to BMD (1973) and not to the more comprehensive BMDP (1975) series.

1. SIMPLE DATA DESCRIPTION AND TABULATION

All three packages can perform various data descriptions and tabulations. SAS and SPSS have the ability to dynamically call any number of routines in a single pass and to produce various descriptions and condensations of the data. BMD does not have this capability; it requires the separate execution of its individual programs each of which is designed for a specific purpose. Therefore, from the point of view of data description and tabulation as well as convenience and economy, SAS and SPSS are better than BMD.

1Part of the results reported herein are from the courses Mathematics 189-678A/679B Advanced Statistical Methods I, II offered by the authors at McGill University during the years 1973-1975.
SAS has separate procedures to print, sort, rank, plot, produce frequency tables (at group intervals, chosen by the user and constructed by program statements) and to compute univariate descriptive statistics of various data sets. Although none of its procedures directly produces a histogram of a data set, this can be accomplished by using appropriate program statements and the procedure PLOT. The above functions can be performed on subgroups of data.

SPSS has the same capability as SAS. The SPSS frequency tables are more comprehensive than those of SAS. It can conveniently produce histograms for discrete (nominal) data (CODEBOOK or MARGINALS in the 1970 version, FREQUENCIES\(^1\) in the 1975 version).

BMD can plot data, produce frequency tables at intervals chosen by the user, compute univariate descriptive statistics, and plot histograms but it does not have separate programs to print, sort or rank the data.

The analogous routines to compute univariate descriptive statistics are BMD01D, SAS/MEANS and SPSS/CONDESCRIPTIVE or CODEBOOK (FREQUENCIES in the 1975 edition). The presentation of the results of BMD01D is generally less appealing than that of its competitors. Moreover, the routines of SAS and SPSS calculate statistics which BMD does not. The distinction for the use of CONDESCRIPTIVE (continuous) and CODEBOOK/FREQUENCIES (discrete) subprograms of SPSS is rather confusing for the lay user. All above routines give the sample mean, standard deviation, minimum and

\(^1\)FREQUENCIES is a combination of the CODEBOOK, FASTMARG and MARGINALS procedures of the 1970 SPSS version.
maximum. In addition, SAS/MEANS gives the sample size, sample total, variance, corrected sum of squares and the coefficient of variation and SPSS/CONDESCRIPTIVE or CODEBOOK give the sample size, variance and the coefficients of skewness and kurtosis. BMDOLD and SPSS/CONDESCRIPTIVE or CODEBOOK give the range and the standard deviation of the sample mean based on a random sample or as otherwise known the standard error. CODEBOOK (FREQUENCIES) of SPSS produces the median and mode of the data. If the routine is used in the general mode the median is not very reliable in view of the particular grouping used. The exact sample median can be obtained for both SAS and SPSS through their data rank routines.

None of the packages can directly produce descriptive statistics or histograms for grouped (tabulated) raw data. This, of course, can easily be accomplished by writing, for instance, a simple Fortran program. Not only do both SAS and SPSS compute more descriptive statistics but their output is presented in an appealing, readable format. Once the sample statistics are available simple statistical inference can be performed routinely.

Multivariate descriptive statistics can be obtained by running multivariate routines as, for instance, the multiple regression routine. SAS and SPSS create their own data set from input data, which, in turn, can be stored for later use, while this is not the case with BMD. Thus the data editing and transforming capabilities of SAS and SPSS are more flexible than BMD.
2. PLOTTING

The PLOT procedure of SAS is one of the best plotting programs available for a line printer at McGill University. Some of its features are: (a) the user may easily specify the exact dimensions of his output; there is no limit to the number of pages of output that may display one continuous plot; (b) axes are automatically printed with X and Y-values given at equal intervals; (c) the user may easily "colour" his graph by specifying that different alphanumeric characters be incorporated for different subsets of (X,Y)-values; (d) "contour-plotting" is possible; (e) any number of graphs may be superimposed; (f) by default or otherwise, if up to n observations fall on the same (X,Y)-value, n different characters (e.g., A, B, C, ...) will be printed correspondingly; (g) any alphanumeric characters may be used.

3. LINEAR REGRESSION

The numerical results of BMD02R, SAS/REGR, and SPSS/REGRESSION agreed for a set of well conditioned simple linear regression data (e.g., a set of data for which the condition number of the centered design matrix was 2.5). On a set of ill-conditioned multiple regression data all three routines calculated the regression coefficients accurately. For the latter set of data SPSS reported a negative error sum of squares whereas the other procedures accurately computed the error sum of squares as zero. SAS/REGR, SPSS/REGRESSION and BMD02R produce all essential statistics needed for the proper evaluation and implementation of linear regression but BMD01R
does not print the intercept of the line, the square of the correlation, or the residuals. Not only do SAS, SPSS and BMD02R compute and print the residuals but they can also plot them. The SAS residual plots are better than the same BMD and SPSS plots. None of the three routines performed a test of lack of fit.

As a general rule, data used in regression analyses, either simple linear or multiple, should be centered. None of the three packages considered here can produce centered data with a single procedure but centering can be achieved using the results of the simple data description subroutines and appropriate transgeneration cards.

SAS through its REGR procedure and BMD02R provide the option to the user for running models with or without an intercept. In the latter case, the computed $R^2$, while it provides a measure of goodness of fit, is not and should not be interpreted as the square of the multiple correlation coefficient.

All routines give the $P$-values (observed levels of significance) associated with various tests in linear regression.

4. STEPWISE REGRESSION

SAS contains two procedures which can select variables in multiple regression. Both of these are easy to use and read. The RSQUARE routine is the only routine (in the three packages under consideration) which is capable of performing all possible regressions. Its output is concise; the complete table is printed only for the equation corresponding to the maximum $R^2$. The STEPWISE procedure of SAS can produce the best regression
equation by utilizing one of five different selection methods. The user has the option of using any number of these in any one run. The user may also specify significance levels for the inclusion or deletion of variables. The default values of these significance levels are 0.5 for inclusion and 0.1 for deletion. The default values for the significance levels for the inclusion or deletion of variables are 0.01 and 0.005 respectively for BMDQ2R and 0.01 for inclusion for SPSS (REGRESSION). Variable levels of selection of variables are also allowed by BMD and SPSS while all possible regressions can be performed with BMD and SPSS by setting up the appropriate regression cards. The output is massive, tedious to read, and no summary table is given. Preliminary cost comparisons on the same system and with the same priority indicate that the SAS and SPSS linear regression routines are cheaper to run than BMD. None of the packages produces Mallows' C_p plots.

5. ANALYSIS OF VARIANCE

Of the more widely used statistical techniques, analysis of variance has a varied treatment in the three packages. BMD has twelve analysis of variance programs, BMDP two, SAS five and SPSS three. One reason for this proliferation of programs is the large variety of analysis of variance designs. In contrast with multiple regression and while the use of the programs is simple and straightforward, the selection of the right routine and the proper interpretation and use of its output requires on the part of the user a deeper knowledge of the intricacies of analysis of variance.
A univariate balanced or unbalanced exact one-way analysis of variance can be performed by BMD04V, BMD07V, BMD07D, BMDP7V, BMDP7D, SAS/ANOVA, SPSS/ONENAY, SPSS/BREAKDOWN, SPSS/ANOVA, BMD05V, BMD06V, BMD10V (X64), SAS/REGR, and SPSS/REGRESSION. The first seven routines are based on sums of squares algorithms while the last six on general linear hypothesis algorithms. With the exception of some refinements, the analysis is the same for both fixed effects and random effects models in either the balanced or unbalanced case. All programs print the analysis of variance summary table and the associated cell means. Additional optional statistics can be printed by all packages. Expected mean squares and estimates of variance components are given only by BMD08V. None of the BMD programs compute the P-level of the associated F-test. These levels, however, are given by the analysis of variance programs of the BMDP series. SPSS is the only package which can compute a homogeneity of variance test through the RANGES option of ONEWAY. Multiple comparisons in one-way analysis of variance can be performed by BMD07V [Duncan's method, Scheffé's S-method, Tukey's multiple range test, the Student-Newman-Keuls procedure, the LSD (least significant difference) method and Bonferroni t-statistics], BMD06V, SPSS/ONEWAY through its RANGES option (the first four methods given above plus the modified LSD procedure and Tukey's alternate procedure), SAS/ANOVA (LSD) and SAS/DUNCAN (Duncan's new multiple range test). The last two SAS procedures require that the error mean square and its degrees of freedom be supplied as input to the program by the user. Therefore, they are less convenient to use than BMD07V or SPSS/ONEWAY with its RANGES option. The description of ANOVA in the SAS manual is very poor and therefore
SAS is not recommended for multiple comparisons. A one-way analysis of covariance with balanced or unbalanced data and one or more covariates can be performed by the sophisticated BMDPIV [comparable to BMD04V (up to 35 covariates) and BMD09V (X82)], BMD01R, and the general linear hypothesis routines BMD05V, BMD06V, BMD10V (X64), SAS/REGR, SPSS/REGRESSION, and SPSS/ANOVA. One-way multivariate analysis of variance and/or covariance can be performed by BMD11V (X63), BMD12V (X69) (balanced data), and SAS/REGR. BMD07M and BMDP7M can also perform multivariate analysis of variance. An exact univariate two-way analysis of variance with balanced data can be performed by the specialized routines BMD02V, BMDP2V, BMD08V, SAS/ANOVA, SPSS/ANOVA or by the general linear model routines BMD05V, BMD06V, BMD10V (X64), BMDP7D, SAS/REGR, and SPSS/REGRESSION. If the data are unbalanced one can use the specialized and versatile BMDP2V, or BMD05V, BMD06V, BMD10V, BMDP7D, SAS/REGR and SPSS/REGRESSION for an exact analysis or SAS/ANOVA, SPSS/ANOVA for an approximate analysis. In the balanced case to perform an analysis with a fixed, random or mixed effects model the user must know which are the appropriate F-tests to be used from the analysis of variance table. The expected mean square column is the criterion to use for the selection of the F-tests. BMD08V is the only program which gives this column. It also gives estimates of the variance components. The program is very easy to use. The easy-to-use SAS/ANOVA procedure does not need the expected mean square column since the tests to be performed can be specified in advance by the user in the TEST statements of the procedure.
In the unbalanced case $BMDF2V$ can handle fixed effects models and split plot designs. No missing values are accepted by the program. No program can perform the Satterthwaite-Cochran approximation in the unbalanced case. Multiple comparisons in two-way analysis of variance can be performed by $SAS/ANCOVA$ (LSD) and $SAS/DUNCAN$ (see comments above in one-way analysis) or by general linear model routines (e.g. $BMD06V$). A two-way analysis of covariance can be performed by $BMD03V$ (balanced data with up to 8 covariates), $BMDP2V$ (balanced or unbalanced data with no specific limit on the number of covariates), $BMD05V$, $BMD06V$, $BMD10V$ (X64), $SAS/REGR$, $SPSS/REGRESSION$, and $SPSS/ANCOVA$. In fact, the latter routine while it cannot perform a full analysis of covariance has provisions for adjusting for up to five covariates in up to a five-way design. Two-way multivariate analysis of variance and/or covariance can be performed by $BMD11V$ (X63), $BMD12V$ (X69) (balanced data), $BMD07M$, $BMDPTM$ and $SAS/REGR$.

An exact univariate three- or higher-way analysis of variance or an analysis of complete factorial experiments or hierarchical designs can be performed by the same routines as in the two-way analysis of variance. $BMD02V$ can handle up to 8 factors of a balanced completely crossed factorial design; it can be used for structures with nesting but the user must be experienced with the art of pooling proper sums of squares.

$BMD08V$ handles up to 9 factors of a balanced nested and/or crossed design with fixed, random, mixed or finite population models. $BMDP2V$ can handle up to 18 factors of a design with complete factorial, crossed or nested, fixed effects structure and balanced or unbalanced data. $SAS/ANCOVA$ handles up to 10 factors of a balanced nested or completely crossed factorial design,
and SPSS/ANOVA up to 5 factors of a balanced or unbalanced factorial design. Analysis of covariance can be performed by the same routines mentioned in connection with two-way analysis of variance. A random effects analysis of variance and/or covariance with a special balanced nested design can also be performed by the NESTED procedure of SAS. SPSS/ANOVA allows pooling of higher-order interactions in the error term, and with its STATISTIC 1 option produces a table for multiple classification analysis (MCA). Multivariate analysis of variance and/or covariance can be performed by the same procedures as in two-way analysis of variance: BMD11V can handle a sizeable number of independent and dependent variables, e.g., up to 55. BMD12V, the multivariate counterpart of BMD08V, can handle only equal cell sizes but it does not use the correct error sum of squares for mixed designs. The program computes the error sum of squares as if the design were fully nested.

SAS/REGR can handle up to 255 variables and SPSS up to 250. SPSS (1973) has no special two- or multi-way analysis of variance procedure. However the 1975 edition of SPSS contains the routine ANOVA designed to cover this case.

SAS/ANOVA allows unequal cell numbers, but except for hierarchical (completely nested) or random effects models the analysis can only be approximately correct. In the unbalanced case the program prints a special warning. The description of the procedure is not clear as to exactly which experimental designs and models it can handle. Besides the designs mentioned above it can handle split-plot designs and, apparently, with expert use of the statements of the procedure, Latin squares,
factorial experiments with confounding, partial replication, etc. No
procedure in the three packages utilizes Yates' algorithm. SAS has a
unique procedure (PLAN) which is capable of generating randomized plans
of experiments.

If the experimental design or data structure is incomplete
(e.g., balanced or partially balanced incomplete block designs, lattice
designs, etc.) the packages offer no specialized routine with the exception
of SAS which offers the procedure LATTICE aimed at analyzing experiments
with one of a certain type of lattice designs. Of course in this case
one can always use a general linear model program or SPSS/ANOVA with its
regression option (OPTION 9).

In general, use of the general linear model routines to perform
analysis of variance is not easy for the inexperienced user. The problem
is that the analysis of variance model is overparameterized and the
original model must usually be reduced to full rank. The analysis of
variance for unbalanced data is not unique and the user of a general linear
model program must determine which analysis of variance is needed, how
to obtain this analysis and how to interpret it.

BMD10V to some extent supplants the BMD05V program. It computes
automatically the values of the dummy variables needed for the analysis
while BMD05V requires that these be supplied as input. In certain
instances this may be an advantage. SAS/REGR, like BMD10V and SPSS/ANOVA,
creates the dummy variables for the classification variables of analysis
of variance. The SAS treatment of missing values is better than EMD. SPSS/ANOVA cannot estimate missing values but can delete cases from subsequent computations. Thus errors with missing cells are only avoided by using SAS. The sequential sums of squares computed by SAS/REGR in analysis of variance are not easy to interpret. An additional feature of SAS/REGR is that it optionally produces a table showing confounding of factors.

Because of the large number of cases of analysis of variance and related techniques and the generally inadequate program description, the inexperienced user is bound to misuse the various programs and their options. For instance in an unbalanced one-way analysis of variance SAS/ANOVA optionally produces a single LSD value instead of a set of LSD values for each pair of mean differences. In other words the programs or manuals do not tell the user whether an option is statistically valid or not. For reasons explained above and from the point of view of the lay user, specialized routines are preferable to the general linear hypothesis procedures. The competing procedures are comparable both in terms of their accuracy and cost.

6. CANONICAL CORRELATION

The routines BMD09M, BMDP6M, SAS/CANCORR, and SPSS/CANCORR perform a canonical correlation analysis with input either the raw data, or a covariance or correlation matrix. Each program computes both the canonical correlations and canonical variables. The outputs of the programs are comparable. EMD09M does not compute any inferential statistics.
It has been supplanted by BMDP6M which computes only the P-values of the inferential statistics. Both SAS/CANCORR and SPSS/CANCORR compute the appropriate chi-square statistic to test the independence of the sets of variates. In addition SPSS gives the Wilks' $\Lambda$ test statistic. The computed values of the canonical correlations are the same for the three packages. Differences exist for the canonical coefficients because different normalizing procedures are employed by the programs. BMDP can handle at least 90 variables, SAS up to 255 and SPSS 250.

7. DISCRIMINANT ANALYSIS

Six routines can perform discriminant analysis: BMD05M, BMD07M, BMDP7M, SAS/DISCRIM, and SPSS/DISCRIMINANT (1973 version) [SPSS/DISCRIMINANT (1975 version)]. The first routine uses no priors (or, in other words, the priors are assumed to be equal). SPSS allows a choice between equal and proportional to sample sizes priors. The other routines offer the user the option to choose and specify his own priors. BMD07M, BMDP7M and SPSS/DISCRIMINANT are stepwise discriminant analysis procedures, i.e., the variables used in computing the linear classification functions are chosen in a stepwise manner. BMDP7M in effect supercedes BMD07M. BMD and SPSS do not employ the same criteria of selection of variables. SAS allows the user to test the homogeneity of within group covariances and then if the groups are not homogeneous to use the individual covariance matrices or if homogeneous to use the pooled covariance matrix thus assuming a common population variance-covariance structure. The 1975 version of SPSS/DISCRIMINANT permits (OPTIONS 14) choice of use of individual or

---

1 This routine is an adaptation of BMD07M.
pooled covariance matrices. SAS can classify a new observation but does not output the classification functions. The opposite is true with SPSS. BMD can do both. They compute the discriminant or classification functions (scores) and thus one can easily classify a new observation. The generalized square distance of SAS is different from the generalized Mahalanobis $D^2$ distance of BMD05M and Mahalanobis distance $D_{ij}$ between groups $i, j$ of SPSS/DISCRIMINANT. In terms of other statistics in the output the programs are comparable. All programs give the classification matrix, i.e., the matrix of classification of sample cases and thus one can evaluate the so-called apparent errors of classification. BMD05M can handle up to 25 variables and up to 5 groups; BMD07M up to 80 variables and has the same limit for groups. SAS can handle up to 255 variables. SPSS can also handle a large number of variables.

8. FACTOR ANALYSIS

Only BMDP4M performs maximum likelihood factor analysis using the Newton-Raphson method of Clarke (1970). This is by far the best program to use. There are five available methods in SPSS/FACTOR but none of these is adequate. Both SAS/FACTOR and SPSS/FACTOR allow various options for rotation and communality estimates. SPSS will also accept a correlation matrix as input. These routines and older BMD routines should only be used if BMDP4M is not available or if it is not possible to obtain a program to perform the maximum likelihood factor analysis as described in Lawley and Maxwell (1971). The old factor analysis routine in BMD was BMD08M.
9. PERIPHERAL DEVICES

The utility and benefits provided by any program or routine are enhanced if the output or selected segments of the computations can be stored in some medium other than paper, e.g., disk, tape or cards. In general and with the exception of the 1975 editions, the three packages BMD, SAS and SPSS possess minimal storage capabilities. They also exhibit wide differences in terms of these capabilities. Before the new editions were introduced, SPSS, for instance, was the only package which could store the correlation matrix (produced by the routines PEARSON CORR, REGRESSION and CANCORR). BMD was also the only package which could save on disk, tape or cards the canonical variables (computed through its BMD09M program). SAS and SPSS (WRITE CASES) can store on disk, drum, or magnetic tape any data set that the user had created. A data set is a collection of input data or data created by means of the SAS or SPSS program statements. The sample means computed by the MEANS procedure of SAS can also be stored. The same is true for the values of all variables, the predicted y's and the residuals used in and/or computed by the procedure REGR of SAS. The procedure WRITE CASES of SPSS may be used to save either the input data or variables which have been computed internally from the original input data.

The 1975 edition of SPSS and BMDP contain major improvements in terms of storage capabilities. Most of the programs of the new SPSS contain additional storage options. A partial list of the routines and their results which can be written on a permanent file is as follows:

FREQUENCIES: all printed output; PEARSON CORR, NONPAR CORR,
PARTIAL CORR, and CANCORR: correlation matrix; REGRESSION: correlation matrix, standardized residuals, standardized estimated y, means and standard deviations; FACTOR: correlation matrix, factor matrix and communalities, factor score coefficient matrix, means and standard deviations; DISCRIMINANT: number of cases in each group, means of variables, pooled within groups correlation matrix, total covariance matrix, discriminant scores, membership probabilities for all groups, actual group and classified group.

Similar improvements exist in BMDP. All of its programs can store the input data on a save disk or tape file. A partial list of the programs and the associated results which can be stored on a save file is as follows:

BMDP1F - two-way contingency tables: observed frequency table; BMDP4M - factor analysis: covariance and correlation matrices, unrotated and rotated factor loadings, factor correlations, factor score coefficients, factor scores, means, variances, sample size and sum of weights; BMDP6M - canonical correlation analysis: covariance and correlation matrices, canonical variables, variances, means, sample size and sum of weights;

BMDP1R - multiple linear regression: covariance and correlation matrices, predicted y's, residuals, variances, means, sample size and sum of weights; BMDP2R - stepwise regression: covariance and correlation matrices, predicted y's and residuals; BMDP6R - partial correlation and multiple regression: covariance and correlation matrices, residuals, variances, means, sample size, sum of weights and partial covariance matrix.
10. GENERAL COMMENTS AND CONCLUSION

The three packages under consideration perform the necessary calculations in double precision unless it is otherwise indicated in the program or stipulated by the McGill computer centre. In any case, except as stated above, the accuracy of computation was satisfactory for most purposes.

The description of the statistical methodology is inadequate in all three packages. This is certainly an area which should receive more emphasis when new user manuals are introduced by the package makers.

*SAS* and *SPSS* have no limit on the number of observations. The file management capabilities of *SAS* are better than those of *SPSS*.

There is some dispute as to which package is "easier" to use. Some users feel that the *SPSS* language is more restrictive and somewhat less natural than *SAS*, whereas others feel that *SPSS* is clearer and more concise. There seems to be some correlation between the first package employed by a user and package preference.

Overall *SAS* and *SPSS* are excellent general purpose non-sophisticated user oriented packages, both potentially better than *BMD*.

ACKNOWLEDGEMENTS

The authors wish to thank Brian Kouri and Michael Edwardes for their contributions and discussions in the course of this investigation. They also wish to thank Professor James W. Frame of U.C.L.A. for his valuable comments on an earlier draft of the paper.
REFERENCES


