SPSS: BEGINNING ITS SECOND DECADE

by

C. Hadlai Hull
Vice President, SPSS Inc.
Chicago, Illinois

Perhaps the best introduction to SPSS--the Statistical Package for the Social Sciences--is to suggest that its name is less than 100 percent accurate. The original intent in 1966 was to develop a system for the analysis of survey data. While the system has grown and now encompasses areas which were outside the scope of the original design, it cannot be said to be primarily a statistically sophisticated system nor can it be said to cater equally to all facets of social science. Let's see what SPSS really is.

First, SPSS is a "system." It is not comprised of individual routines for statistical analysis but is a single conceptually and physically integrated system. Second, SPSS is a system for "analysis." While the emphasis, in terms of output, is on statistical measures of the user's data, the system is intended for analysis in a more general sense and offers the user the means of modifying the data and displaying the data other than through statistical measures. Finally, SPSS was originally intended for the analysis of "survey data" in particular. Thus, it was not designed to cater equally to the needs of experimentalists or those analyzing aggregate data.

What are the identifying features of the system which result from its orientation toward the analysis of survey data? It is equipped to handle relatively large numbers of subjects--particularly in the range of two to five thousand. Second, it can deal with large numbers of variables--up to one thousand at the present time. These variables are, in turn, known to be largely nominal in nature and are frequently missing for a subset of the population. The type of data and the classical methods of analyzing this type of data led to emphasis on analytic procedures such as frequency and crosstabulation displays. Finally, the statistical and data processing expertise of the potential user of such a system was judged to be relatively low, thus a great deal of emphasis was placed on developing a system which, with its documentation, would place survey analysis within easy reach of this category of user.

The fact that the current SPSS batch system is in use at roughly one thousand computer installations spread through all fifty states and some fifty foreign countries and operating on virtually all major hardware systems suggests that it fulfills at least some requirements quite well. On the basis of feedback from the user community, future developments should be aimed toward fields of research other than survey analysis and toward conversational analysis. Such developments are currently underway, and I am able to describe not just the SPSS batch system with which most university computer centers are acquainted, but also the nearly-complete SPSS conversational system which we hope will be generally available at this time next year.

The SPSS batch system is perhaps best known for its ease of use by those who are unsophisticated in the area of data processing. The system specifically has a simple, user oriented, free field control language. While the control language may lack sophistication in some respects, the fact that each statement accomplishes one straightforward action makes it simpler to explain and comprehend. Of major consequence in making the system simpler to use is the existence of the "self-defining" file--an SPSS specific data file format which includes not only the data

themselves, but also the machine-readable description of the data such that the user need describe the data only once with subsequent analyses taking advantage of the original definition. While the use of such self-defining files has become widespread, SPSS was one of the earlier systems to incorporate this feature.

An ongoing effort has been made to keep the printed output generated by SPSS cleanly laid out and intelligible to even the statistically unsophisticated. It specifically includes both the user-defined mnemonic names of the variables and descriptive labels for the variables and the categories of nominal variables. The intelligibility of the output is accompanied by what our users feel is perhaps the major strength of the system--a complete and highly intelligible manual written by a social scientist for social scientists.

Although the original intent of the batch system is still apparent, the current system includes sixteen different analytic procedures, which can be categorized as follows:

The "old standbys" of SPSS are the frequency display routines--in particular, the FREQUENCIES routine and the CROSSTAB routine which provide one-way and n-way frequency displays respectively. The next release of the batch system will offer yet another routine in this area--a routine designed to display one-way and n-way frequency tables for multiple-response data. Multiple-response data result from interview questions such as "Which news magazines do you read?" which elicit more than one response.

A second general area of analytic technique comprises non-parametric statistics. This area includes the existing Spearman/Kendall rank-order correlation routine and will, in the next release, include a routine which will perform virtually all the standard non-parametric tests such as the Kruskal-Wallis H test and the Cochran's Q test.

A third area is analysis of variance. The existing routines in this area are one-way and n-way analysis of variance. The latter, however, falls short of the requirements of certain areas of research--particularly experimental research--and a new analysis of variance routine has been developed. This new routine offers multivariate analysis of variance and gives the user virtually complete control over the specification of the design to be analyzed.

The fourth general area of analytic routine includes those which are based in one way or another on product moment correlation coefficients. The more frequently-used routines in this area are the correlation routine, partial correlations, regression and factor analysis.

Finally, there are several analytic routines which do not fit into any of the above conceptual categories such as T tests, discriminant analysis, and Guttman scaling.

I emphasized earlier that the SPSS system includes not just statistical routines, but also data and file modification capabilities. In particular, the user may add or delete both variables and observations from an existing SPSS file. New variables may be either read in or calculated within SPSS on the basis of existing variables. In addition, existing variables may be modified by means of the SPSS recode capability. The file modification capabilities include both introducing new cases in machine-readable form and taking a random sample of a file, weighting the cases in a file, and selecting cases based on their contents.

Installations having a maintenance contract with SPSS receive, once every twelve to eighteen months, an enhanced version of the batch system. These new versions typically include two or three complete new analytic routines, several major enhancements, and improvements to existing procedures and data-handling facilities as well as correction of all known errors. Further, installations receive periodic intermediate revisions to effect enhancements and corrections, distribution of information regarding bugs, and telephone consultation with our staff in Chicago. Of course, installations with other than IBM hardware would contact the site providing the specific conversion.

I wouldn't be honest if I didn't mention that the current SPSS batch system has some shortcomings. First, it does not appeal equally to all segments of the potential social science user population. In particular, we are aware that the system does not give the econometrician the wealth of regression techniques available elsewhere. Also, as noted previously, the ANOVA routine falls short of the requirements of the experimentalist. You might also feel that the lack of AID or a cluster analysis routine is detrimental if you are attempting to service all your users with a single statistical system. However, without neglecting

the survey analysts who have been our primary audience, we attempt with each release to broaden the scope of the SPSS batch system and to upgrade the statistical and numerical algorithms in existing facilities. Shortcomings in other than the statistical areas include the fact that the batch system is equipped to deal only with rectangular data files--that is, all cases must represent the same unit of analysis and must include the same variables. Finally, the system is a single very large program and cannot, in general, be adapted to the mini-computers such as the PDP 11 which are becoming so popular on campus.

The popularity of the SPSS batch system has generated sufficient income for us to devote significant resources to the development of a totally new product--the SPSS conversational system. This system is not, unfortunately, currently available for distribution. However, it is nearing completion and should be available to be run under IBM's Time Sharing Option (TSO) in the near future, and will probably be adapted to run on the DEC 10 and under the Control Data KRONOS system during the next calendar year.

The SPSS conversational system is not an adaptation of the batch system running in a conversational environment. It is a totally new system designed from the bottom up to be truly conversational. Its command structure and display facilities are both oriented toward the terminal user. In particular, the system will lead the user through the steps required for definition of a problem with successive prompts or will permit the user to lead the system by specifying both the next prompt and the response. It features different modes of prompting which are adapted to different levels of user expertise and has the capability of responding to a command of HELP with information on the options available to the user under various situations. It is intended for use from hardcopy and CRT keyboard terminals with line widths of sixty characters and up and from twelve lines per screen on up. It also has the facility of generating a machine-readable record of a terminal session for subsequent listing on a high speed printer.

It is our feeling, as borne out in classroom use of the system during the development phase, that conversational data analysis is very different from the analysis currently done in the batch. Conversational analysis permits the user to focus on one individual aspect of a research problem and to refine the hypothesis in an iterative fashion. This process is relatively conservative of computing resources and produces a modest amount of printed output since the researcher is not encouraged to investigate multiple potential relationships during a single computer run as in the CROSSTAB ALL BY ALL case. I think that one can postulate that holding output speed down to thirty characters per second can do a great deal to promote better research methodology.

I think that you will find that the new conversational system is oriented to much the same audience as the batch system. It features readable display output complete with mnemonic variable names and variable and category labels. The four statistical techniques currently available are roughly the equivalent of the batch FREQUENCIES, CROSSTABS, BREAKDOWN and PEARSON CORR procedures with partial correlation available as a by-product of Pearson correlations. The system appears to offer reasonable response time, with the response time being more a function of the overall system load than what the terminal user asks for, although response time will suffer as the number of cases gets larger. The analysis of the typical two thousand subject survey file does not appear to be any problem, and, due to the organization of the file in the conversational system, response time is independent of the number of variables in the file.

The conversational system is designed to accept a more or less unlimited number of variables with the availability of on-line disk space more likely to be the limiting factor than the capacity of the system. The system currently permits somewhat more than 750 variables, but the architecture is amenable to permitting five thousand or more variables in future releases.

While the conversational system provides its own mechanism for defining and performing the equivalent of the batch "SAVE FILE" in order to generate a self-defining file for subsequent analysis, we have also provided the means for creating a conversational file with the batch system. Thus existing SPSS data bases, other than those containing variables with non-numeric codes, are immediately available for conversational analysis.

We think that an aspect of the conversational system which is nearly as valuable as the self-defining file was to the batch system is the "workfile" concept. Essentially, conversational analysis is conducted using two data files. One is the "master" file, which contains pretty

much the same information as contained in a batch system file. The other file is the "work" file. A work file is associated with a specific master file and contains only those aspects of the master file which the user has modified through recoding, transformations, weighting, case selection, etc. The advantages of this concept are as follows: First, several users, be they students or members of a research team, can analyze the same master file without getting in one another's way or needlessly duplicating data. Second, the master file itself is never changed, so that inadvertent destruction of data is made more difficult. Third, although we have attempted to code the system so that it is not subject to loss of file integrity in case of a system crash, the fact that the master file is never altered provides additional security. There is, of course, a facility for merging a work file with its master file to produce a new master file after the researcher has determined which new variables, etc., are worth retaining.

In addition to file definition and the four statistical routines, the conversational system currently includes a facility very much like the IF and COMPUTE statements of the batch but with some extensions such as a LAG function. It also provides the ability to weight the cases in a file and to select cases based on their contents.

The immediate future should see the development of those additional facilities which we feel are required prior to general release of the system. Specifically, the system needs the equivalent of the batch RECODE facility and a REGRESSION procedure. The latter has already been designed with a great deal of attention given to the facilities which our users tell us are missing in the batch regression procedure--particularly backward elimination and a less awkward method for attaching residuals to a file. Other required facilities are the naming, saving and retrieving of correlation matrices and an initial cut at documentation. We frankly do not know what form the documentation should take for the conversational system since the system itself is largely self-documenting. However, at least primitive documentation aimed at the user who knows what he wants to do will be available with the initial release.

The most serious limitations of the conversational system which are apparent today are that the system was not designed primarily for performance and that there is a fixed limit to the number of "active" variables, i.e. no statistical procedure may analyze more than twenty variables concurrently. This active variable limitation is likely to prove burdensome in the areas of regression and factor analysis and may well be eliminated. The significance of the fact that the system was designed primarily to give the serious researcher a complete facility rather than to give maximum performance with limited function is that the conversational system may prove expensive to use in large introductory courses and will probably be quite difficult to adapt to hardware of modest capability.

While we have our hands full trying to stay abreast of the requirements for further extensions to the batch system and the necessity of wrapping up the development of the initial release of the conversational system, we have given some thought to where the conversational system should go after the first release. At this point, it would appear that our priorities should be on the incorporation of some additional procedures such as factor analysis and analysis of variance and on the handling of other than traditional rectangular files. In particular, we have designed the system with the handling of hierarchical files in mind. Typical hierarchical files are the U.S. Census Public Use Samples which include records describing different units of analysis such as political unit, household, and individual in a single file.