SPSS/HP
Statistical Package
For The Social Science
Hewlett-Packard Version

An Update

Marlys A. Nelson
Western Wisconsin Academic Computing Consortium
University of Wisconsin-River Falls
River Falls, Wisconsin 54022

Nicholas Elliott
Political Science Department
University of Wisconsin-River Falls
River Falls, Wisconsin 54022

SPSS/HP, an acronym meaning "Statistical Package for the Social Sciences"/Hewlett-Packard version, encompasses a variety of routines whose function it is to perform statistical analysis of large amounts of data. Also included are easy data entry, selection and modification facilities.

Development of SPSS/HP began in 1974 at DePaul University in Chicago, Illinois for the HP 2000 C'/F time-sharing system. This development was begun due to a growing need by their users for a meaningful statistical analysis system capable of being run on a small time-sharing system. The result was SPSS/HP, an authorized version of SPSS, the well-known package of statistical programs commonly used for research and instruction in the social sciences.

Two years later, the University of Wisconsin-River Falls became interested in the package and converted SPSS/HP to the HP 3000 computer system, expanding upon the HP2000 version to make efficient use of HP3000 system resources. Both sites are continuing to improve and enhance the package.

In designing such a system, the aim was to develop a package which would be familiar to experienced SPSS users but would also take full advantage of the time-sharing features and resources of the HP systems. A user who has gained experience with SPSS at another institution and/or involving another computer system can, with little effort, use the SPSS/HP package.

SPSS/HP is a fully interactive system. All a potential user is required to know is the proper log on. More sophisticated SPSS/HP users may take advantage of HP3000 features, such as the STREAM facility to operate SPSS/HP in a batch mode.

There are four (4) major areas in statistical analysis with which SPSS/HP is concerned. These areas are:

1. Data entry,
2. Data modification,
3. Data selection, and
4. Statistical analysis.

One of the more time consuming, if not tedious, tasks involved in statistical analysis is that of data entry. In an attempt to make this task less burdensome, several methods have been incorporated into SPSS/HP to accomodate several possible data forms and methods of data entry. In the most straightforward, simplest method, data may be entered directly into a SPSS/HP disc file from a terminal keyboard. Such an example is shown in Appendix A. Alternate data entry methods include reading from magnetic tape, punched cards, optical mark cards or existing disc files. Data may also be entered in several forms, such as the traditional 80 'column' card format, BASIC data files (BASD type) or the traditional keyboard entry where each data element is separated by a comma. All these data entry options were designed to give the researcher flexibility in setting up his analysis.

Once all the necessary data has been entered, the user may begin to work with the data. Several facilities are included in SPSS/HP to allow data editing to correct erroneously entered data values, the modification of existing data and the creation of new data through standard mathematical operations. Existing data may be modified by performing an arithmetic operation, such as adding 10 to all data values for one variable, or by editing values individually, such as adding two existing variables, or by using one of the several SPSS/HP functions. The functions include routines which will create various statistical distributions, such as NORMAL, CHI, T or F distributions. Also, new data items may be created through the use of a mathematical-type formula of the form 'NEWVAR=SQRT(4*VAR1)'.

Facilities have also been included to allow the user to select for analysis at a given time only a subset of the entire data file. This selection may be done by selecting only cases where one of the variables involved takes on a specified value. For example, suppose a variable called SEX is coded to indicate 1=MALE respondent and 2=FEMALE respondent. In one specific instance, let us assume we wish to analyze only the MALE responses. This could be accomplished by selecting for analysis only the cases where the variable SEX has the value 1. Subgroups of the total data file may be devised in such ways.

The last step is then the statistical analysis itself. SPSS/HP currently includes 15 different statistical routines. See Appendix A for a list and brief description of the function of each. Also included in Appendix A is a sample run of the newest statistic, ANOVAU, which performs an analysis of variance for designs which have an unequal number of subjects per cell.

In developing, converting and expanding a package of this nature, problems do occur which may -- depending upon the method undertaken to correct the problem -- affect the overall design of the package. Three such problems encountered during the work on SPSS/HP for the HP3000 will be described below. Hopefully, this will serve a dual purpose. First, to illustrate some of the considerations dealt with specifically when SPSS/HP was developed, and secondly, these same problems and solutions may apply to any type of work in the area of statistics or the conversion of large program packages.

One major concern an interactive program package must face is that of execution speed. A user does not want to sit in front of a terminal keyboard for endless periods of time waiting for calculations to be performed. Ideally, the user hopes to press a few keys and have the answer flash on his screen. He can then examine the results and decide upon his next course of action. Two programming concerns contribute towards execution speed -- the amount of time spent reading and manipulating the data file to obtain the necessary information and the size of each program segment.

The first factor, the amount of time necessary to read and manipulate the data file, is obviously best if minimized. This

idea has been followed in SPSS/HP where, with minor exceptions, the data file is never read more than once for any given statistic. In a few instances, this concept was not followed in the attempt to allow the program to determine from the data several factors which otherwise the user would be required to supply. One such example is the routine which performs an analysis of variance for balanced designs, ANOVA. SPSS/HP requires two passes over the data file in this instance. The first pass identifies the specifications of the design, such as the number of levels within each factor in the design and the number of subjects per cell in the design, while the second pass reads the data to perform the actual statistical calculations. If only one pass had been made over the data, the ANOVA user would have been required to supply the routine with the specifications of the design. By compromising in this instance on the number of times the data file is read, the use of ANOVA was simplified.

The second factor playing a role in determining program execution time, the size of each program segment, proved to be a double-edged sword. Again, ideally, the smaller a program segment, the faster execution time obtained. The computer is able to spend less time swapping program segments in and out of main memory since smaller segments force less displacement of other program segments already memory resident. Striving for a small program segment size creates another problem. The actual number of program segments is then increased, one large segment split into two smaller segments for example. The actual number of program segments in any one program has a finite upper limit of 63, however. Faced with such a problem, it appears that the only solution lies in a larger program segment size with the hope that execution time does not degrade to a substantial degree. Upon closer examination, however, another solution becomes apparent. While it is true each program may contain at most 63 program segments, there seems to be an unlimited number of programs which may be linked together thru the concept of Process Handling. In process handling, program A, termed the father process, begins execution. When condition 1 occurs, it is directed to activate program B1 for execution. Program B1 becomes the son process and runs to completion at which time program A, the father, again begins execution at the point following the request to activate and execute the son. This linkage of programs can be extended further whereby the father process may activate different son processes and each son may become a father and activate another son. By using this technique, response time did not differ drastically and the various segments within the SPSS/HP package can remain small -- yet the total package remains highly expandable.

A major concern faced when dealing with any mathematical formula to be calculated by computer is the problem of precision. Currently, standard precision of six (6) digits on internal computations is employed. Output values may often be less accurate, however, such as four (4) digit accuracy. Extended precision, type LONG in BASIC, would provide accuracy up to fourteen (14) digits but was not implemented for several reasons. The initial data is entered in standard precision. The entry of data accurate to fourteen (14) digits is rather ridiculous. The need for precision arises in the calculations required for a given statistic.

In this case, a statistical routine would need to read the data in standard precision and convert it to double precision for use in calculations. In an attempt to do this on a limited scale in routines such as REGRESSION, it was found that internal BASIC system routines controlling the conversion and formatting of numeric output could not handle such tasks precisely. Beyond the 4th or 5th digit, accuracy was lost which meant the conversion to LONG served no purpose. An alternative method would be the storage of all data in LONG form initially, eliminating the need for conversion; but this would double the storage necessary for all data files, an unhappy thought for systems constrained by on-line disc storage. At present, it was felt that all calculations were precise enough to delay any efforts at further solutions.

The final problem encountered evolved slowly. One basic design goal set at DePaul University was to avoid the use of scratch disc files. Some programming technique, similar to using scratch files, was necessary, however, to handle situations in which data was deleted or added from an existing data file. The use of the original data file twice within a routine was developed as a solution to this problem. The original data file was assigned as two separate files within a given routine, for example, as files #1 and #2 in a BASIC program. The old information would be read from file #1, the addition or deletion or modification of the data would be made; and then the information would be printed on file #2. This was a nice technique for the HP2000, which can handle the file buffering to avoid reading on file #1 the information just printed on file #2. It was discovered, however, that the HP3000 was not designed to handle such file manipulation. Overlapping of the file information began to occur after a certain number of file reads and prints. Faced with the failure of this technique, the only feasible alternative was the use of scratch disc files, a task easier on the HP3000 than on the HP2000. Temporary disc files may be built and used without any user being aware of the existence of such files. The only restriction imposed by the use of scratch files is that sufficient disc space must exist so that a temporary file identical in size to the original data file may be created---a restriction which has not shown itself to be a hardship as of yet.

Several enhancements of SPSS/HP have been developed in the past few months. The first such enhancement increased the maximum number of variables which may be contained in an SPSS/HP data file. Previous versions have had an upper limit of 108 variables, but this limit has now been raised to allow 500 variables. As before, the number of cases is limited only by the amount of available disc space.

A second recent enhancement is the capability to enter variable labels. This feature allows each variable declared on a SPSS/HP data file to be given a 1 to 40 character description intended to explain the purpose of the variable. These variable labels may be added, edited, or deleted at any time. Variable labels will be printed by each statistical routine. In the future, it is planned to include value labels, which would allow up to 10 values per variable to be given a 1 to 20 character description. This description would also be printed by each statistical routine.

The last major enhancement to SPSS/HP has been the addition of a new statistical routine. This routine, ANOVAU, performs a fixed effects factorial analysis of variance for a design in which there exists an unequal number of subjects per cell. ANOVAU allows for either a two-way or three-way analysis of variance and no factor within the design may contain more than 40 levels. The routine provides the experimentor with four (4) models with which he may test his hypothesis. These models are: 1) The complete linear model; 2) the method of fitting constants; 3) hierarchical analysis; and 4) the unadjusted main effects analysis. A complete table of cell means and variances is also printed along with the analysis of variance summary table. ANOVAU will also handle a design which may contain up to five (5) covariates.

APPENDIX A

Included in the appendix is a sample run showing data entry from a terminal keyboard into a SPSS/HP data file.

Also included is a table listing all the statistical routines currently available through SPSS/HP and a brief description of the function of each of these routines.

The output produced by the newest statistical routine, ANOVAU, is enclosed. This serves to show the form of the statistical output and to illustrate the routine itself.

```
:RUN SPSSHP


SPSSHP : VERSION 4.04          5/31/78
SPSSHP NEWS LAST CHANGED ON    5/31/78
WED   OCT  4, 1978     8:16 AM


TYPE NEWS FOR NEW FEATURES AND KNOWN PROBLEMS IN SPSSHP.
TYPE INSTRUCT FOR BASIC DIRECTIONS FOR USING SPSSHP.


NEXT ? CREATE FILE!DATA00,3,24,LABELED

FILE SUCCESSFULLY CREATED - USE 'FILE NAME!DATA00' TO WORK
WITH THIS NEW FILE.


NEXT ? FILE NAME!DATA00

NEXT ? VARIABLE LIST!READ,YEAR,ACTIV

NEXT ? N OF CASES!24

NEXT ? READ INPUT DATA

ENTER DATA ONE CASE AT A TIME.
TYPE 'PROMPT' OR 'NOPROMPT' TO TURN PROMPTS ON OR OFF RESPECTIVELY.
'STOP' WILL INTERRUPT DATA ENTRY.

CASE    1, READ TO ACTIV
?1,1,8
CASE    2, READ TO ACTIV
?1,2,5
CASE    3, READ TO ACTIV
?1,2,8
CASE    4, READ TO ACTIV
?1,3,2
CASE    5, READ TO ACTIV
?1,3,7
CASE    6, READ TO ACTIV
?1,3,7
CASE    7, READ TO ACTIV
?1,3,9
CASE    8, READ TO ACTIV
?1,4,5
CASE    9, READ TO ACTIV
?1,4,7
CASE   10, READ TO ACTIV
?1,4,7
   .
   .
   .
CASE   24, READ TO ACTIV
?2,4,2

DATA ENTRY COMPLETED.
```

STATISTIC NAME	FUNCTION
CONDESCRIPTIVE	Computes ten descriptive statistics for interval level data, including the mean, variance and standard deviation.
FREQUENCIES	Computes frequency distributions for categorized data. Histograms may be requested.
BREAKDOWN	Computes descriptive statistics for subgroups of cases. A one-way analysis of variance can also be performed.
CROSSTABS	Computes contingency tables from categorized data. Four non-parametric statistics may be requested.
T-TEST	Computes the t statistic employing one of three user defined parameters of difference of means.
PEARSON CORR	Computes Pearson product-moment correlations.
SCATTERGRAM	Performs bivariate regression and prints scatter diagrams.
REGRESSION	Performs multiple regression. A backward stepwise procedure is available.
RESIDUAL	Performs residual analysis. Durban-Watson statistics and a plot of autocorrelation functions of residuals and residuals against time may be plotted.
ANOVA	Performs a factorial analysis of variance for balanced designs.
ANOVAU	Performs a fixed effects factorial analysis of variance for unbalanced designs.
DISTRIBUTIONS	Computes frequency distribution for continuous data. Histogram option is available.
SPEARMAN CORR	Computes Spearman's rank-order correlations.
T-SQUARE	Computes Hotelling's T(2) statistic
U-TEST	Performs Mann-Whitnet tests.

```
NEXT ? ANOVA(1)ACTIV BY READ,YEAR

NEXT ? OPTIONS(3)

NEXT ? EXECUTE

STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES - NF VERSION 4.04

FILE : DATA00

(CREATION DATE = WED, OCT  4, 1978,  8:18 AM )
```

A N A L Y S I S O F V A R I A N C E

	ACTIV	STUDENT ACTIVISM
BY	READ	STUDENT READING LEVEL
	YEAR	YEAR IN SCHOOL

MEAN AND VARIANCE TABLE

VARIABLES	READ	YEAR	MEAN	VARIANCE
FOR FACTOR : READ				
LEVEL	1.00	.00	6.500	4.056
LEVEL	2.00	.00	3.357	2.863
FOR FACTOR : YEAR				
LEVEL	.00	1.00	4.333	6.667
LEVEL	.00	2.00	4.600	7.300
LEVEL	.00	3.00	4.875	6.696
LEVEL	.00	4.00	4.800	5.200
FOR FACTOR : READ X YEAR				
CELL	1.00	1.00	8.000	.000
CELL	1.00	2.00	6.500	4.500
CELL	1.00	3.00	6.250	8.917
CELL	1.00	4.00	6.333	1.333
CELL	2.00	1.00	3.600	4.300
CELL	2.00	2.00	3.333	6.333
CELL	2.00	3.00	3.500	1.667
CELL	2.00	4.00	2.500	.500

GRAND MEAN = 4.667

ANALYSIS OF VARIANCE : METHOD OF FITTING CONSTANTS

SUMMARY TABLE

SOURCE OF VARIANCE	SUM OF SQUARES	DEGREES FREEDOM	MEAN SQUARES	F	PROB. OF F
MAIN EFFECTS	60.091	4			
READ	58.966	1	58.966	13.62	.0000
YEAR	2.472	3	.824	.19	.9054
TWO-WAY INTERACTIONS	1.959	3			
READ X YEAR	1.959	3	.653	.15	.9908
ERROR	69.283	16	4.330		
TOTAL	131.333	23			

NEXT ? FINISH

RUN ENDED