SAS COMMUNICATIONS

INSTITUTE OF STATISTICS, NORTH CAROLINA STATE UNIVERSITY, RALEIGH, NORTH CAROLINA #1 DECEMBER,1973

The first versions of the Statistical Analysis System remained pretty much within the domain of North Carolina State University. In the sixties, a typical SAS user worked at N.C.S.U. and had only to call or walk across campus to discuss the system with its developers. The user had full opportunity to find out about the then-current capabilities of SAS, the plans that Messrs. Barr and Goodnight had for the system, some of the out-of-the-ordinary tasks to which SAS could be put, and so forth.

Expanded and refined into what we believe to be the most powerful, flexible system extant for data management and analysis, SAS now boasts users at more than 100 installations across the United States and overseas. The informal network of communication is no longer adequate for serving the enlarged community of SAS users. We of the SAS project group are therefore launching SAS Communications. This newsletter will, we hope, involve users in the continuing development of SAS and aid them in their use of the current version. SAS Communications will be published at irregular intervals in an informal format, for we value the contributions that easy, spontaneous interaction between users and developers have already made and assuredly will make to the system.

The newsletter will highlight SAS features that are commonly overlooked. It will outline the activities of the SAS project group and the goals that the group intends to meet. Application notes -- special ways of using SAS -- will be featured too, as will notes on the algorithms embodied in SAS. We shall describe additions to the Supplementary Procedures Library, the collection of special-purpose procedures written by the staff and by users. Profiles of the people who make up the project group will be included. Other publications about SAS will be listed. We also plan to initiate a users' forum, wherein a user can question or comment to the staff and other users. Finally, every few issues, we will include an up-to-date index of the articles in SAS Communications.

We urge you, the users, to contribute to SAS Communications; it is only with your help that SAS can grow to meet the needs of the community it serves.

Jolayne Service

for the

SAS Project Group:

A. J. Barr, Systems
J. H. Goodnight, Procedures and Administration
J. Service, Documentation
C. G. Perkins, Assistant for Systems
J. Sall, Assistant for Procedures
B. Reeves, Administrative Assistant
H. J. Kirk, Consulting
S. L. Biggs, Consulting

SAS GROWS

A. J. Barr is currently rewriting the "systems" portion of SAS, that part which builds data sets and processes SAS statements. Multiple aims characterize his activity.

First, Mr. Barr is making it easier to expand the capabilities of the system. He is also making the internal workings of SAS clearer to other systems programmers, who might wish to make special-purpose modifications of SAS.

The rewritten system will handle data sets better, especially partitioned data sets. Less Job Control Language will have to be used to store and retrieve SAS data sets. Data sets on magnetic tape will acquire some of the flexibility now associated with disk data sets.

An installation will be able to "autobatch" SAS jobs; that is, to run series of small SAS jobs end-to-end very efficiently. Users familiar with autobatched versions of the programming language compilers WATFIV and PL/C will recognize the advantages of autobatching. Also, SAS will be easier to invoke interactively under the Time Sharing Option (TSO) of the IBM operating system.

The new SAS will include report-writing capabilities. A user will be able to make SAS produce a customized, possibly annotated report, easily readable by people unfamiliar with computers.

Finally, Mr. Barr is making SAS statements more consistent and flexible. For example, where a user would now write

```
Q2 Q3 Q4 Q5 Q6 Q7
```

he will be able simply to write

```
Q2-Q7
```

Other members of the SAS project group are developing and modifying SAS Procedures. J. H. Goodnight is preparing a procedure to perform non-linear least-squares curve-fitting. Expanded capabilities are being added to DISCRIM, the discriminant analysis procedure. Carroll Perkins is working on a procedure for producing histograms.

NEW SPONSORS FOR SAS

The Directors of the Southern Regional Agricultural Experiment Stations approved in June, 1973, the project titled, "The Statistical Analysis System: Its Development and Maintenance." This project provides a new source of funding for SAS, insuring its growth for the next five years. Participating in the project are the State Agricultural Experiment Stations of Alabama (at Auburn University), Florida (at the University of Florida),Georgia (at the University of Georgia), Kentucky (at the University of Kentucky), Louisiana (at Louisiana State University), North Carolina (at North Carolina State University), Oklahoma (at Oklahoma State University), South Carolina (at Clemson University), Tennessee (at the University of Tennessee), Texas (at Texas A & M University) and Virginia (at Virginia Polytechnic Institute and State University), and the Agricultural Research Services of the United States Department of Agriculture. A Technical Committee was formed, which includes a representative from each participating organization and has Clyde Y. Cramer of Virginia as chairman, Robert D. Morrision of Oklahoma as vice-chairman, and Robert J. Monroe of North Carolina as secretary.

The subcommittee on price structure has set yearly fees for SAS installations Details about the charges and the services to which those who pay are entitled are available from J. H. Goodnight, Institute of Statistics, Raleigh, N. C. 27607.

The initial development of SAS was supported in part by grant RR-00011 from the Biotechnology Resources Branch, Division of Research Resources, NIH.

SUPPLEMENTARY PROCEDURES

This fall, we released the SAS supplementary procedure library, a collection of special-purpose routines. Some were contributed by SAS users; all are maintained by the project group. For those who have not yet seen the supplementary procedures guide, we list below the procedures presently documented.

PRTPCH: For punching or printing SAS data sets in a format determined by a user-written FORTRAN IV specification.

INBREED: For calculating inbreeding or covariance coefficients.

HARVEY: A SAS implementation of Dr. Walter Harvey's Least Squares and Maximum Likelihood General Purpose Program.

RENAME: For changing the names of variables in a SAS data set.

QUESTN: A quick procedure for producing frequency and contingency tables for data from questionnaires.

EXPLODE: For printing characters over one inch high.

PROBIT: For probit analysis of biological assay data.

STANDARD: For standardizing the values of numeric variables to a given mean and standard deviation.

A NOTE ON PRECISION

Some users have asked how precise are the results SAS produces. The answer, of course, depends somewhat on the data submitted to SAS and on the procedures used. We note, however, that all numeric data is stored and manipulated in double precision--about 16 decimal digits.

In 1967, James W. Longley published in the Journal of the American Statistical Association an article entitled "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User." It presented some data for which regression analyses were to be performed and the results of using several then-popular computer programs for doing the calculations. He found that the error in the calculation of one typical regression coefficient ranged from .03% (in the most accurate program considered) to 375%.

This year, Terry Seaks of the University of North Carolina at Greensboro used SAS and several other newer programs for analyzing the Longley data. SAS virtually tied with another system for lowest error rate. Its answer for the regression coefficient mentioned was accurate to eight digits.

SAS PUBLICATIONS

Service, Jolayne. A User's Guide to the Statistical Analysis System. 1972.

Perkins, Carroll Gray. A Guide to the Supplementary Procedure Library for the Statistical Analysis System. 1973.

Barr, Anthony James, and James Howard Goodnight. SAS Programmer's Guide. 1972.

The user's guide, published by and available from the Student Supply Stores, North Carolina State University, Raleigh, North Carolina 27607, tells how to use SAS. The supplementary procedures guide describes special-purpose SAS procedures. Programmers who want to implement their own procedures under the umbrella of SAS will need to consult the programmer's guide. The latter two guides are available from the SAS project group, Institute of Statistics, North Carolina State University, Raleigh, North Carolina 27607.

ADDENDA TO THE USER'S GUIDE

A few facilities of the current version of SAS were omitted from the User's Guide. We circulated to all installations a memorandum detailing the omissions; here, we note those items for users who missed the memorandum.

Changes to the INPUT statement description

1. An INPUT statement can introduce packed decimal data into SAS. If the values of the variable AGE were in packed decimal form in positions 3 through 6 of the input records, one would write

   ```
   INPUT AGE PD 3-6;
   ```

2. The maximum number of characters per observation is 32000.

3. The number of cards per observation is not restricted.

4. Values of character variables may be no more than 80 characters long.

Changes to the descriptions of program statements

1. A KEEP statement of the form

   ```
   KEEP variable_1 < variable_2 ... variable_n>;
   ```

   can be used to retain only the listed variables in the data set being built.

2. In building a SAS data set, a numeric variable named ERRORSW is generated automatically but not included in the observations actually added to the data set. ERRORSW is set to zero before an observation is constructed; it is set to one if an invalid data element or an arithmetic error (like an invalid function argument) is encountered in forming that observation.

3. A Boolean operator YES can be used in place of NOT NO.

4. Two comparision operators can surround a quantity. For example,

   ```
   AGE > 12 AND AGE < 20
   ```

   can be abbreviated

   ```
   12 < AGE < 20
   ```

5. A STOP statment, written

   ```
   STOP;
   ```

   forces SAS to ignore the observation being processed and to cease altogether building the data set.

6. When an observation is to be formed from data on several records, one can use a LOSTCARD statement to insure that all "lost" records are detected. The statement terminates the formation of an observation and tells the INPUT statement to return to the records it just read, ignoring the first record and drawing in the appropriate number of records beginning with the second record previously encountered. For example, one might write

   ```
   INPUT ID 1-3 S1 5-10 IDCHECK #2 1-3 S2 #2 5-10;
   IF ID¬=IDCHECK THEN ERROR ID IDCHECK;
   IF ID¬=IDCHECK THEN LOSTCARD;
   ```

Other Changes

1. Unless DUMMYB is specified in the MODEL statement in the REGR procedure, the rows and columns associated with dummy variables in the XPX, SWEPT, and INVERSE matrices will not be printed.

2. A new statement for use under TSO,

   RUN;

   forces SAS to execute all statements already entered but not executed. Additional SAS statements can be entered after the results of the execution are printed.

APPLICATION NOTE: RANDOMLY SAMPLING OBSERVATIONS IN A SAS DATA SET

Occasionally when a statistician must consider a very large data base, he finds it useful to investigate just a sample of the observations recorded. SAS's ability to generate pseudo-random numbers makes it easy to use the system to isolate a random sample.

To take a simple random sample of the observations in a data collection, the SAS user could create a variable which associates a pseudo-random number with each observation. The data set would be sorted according to those random numbers, thus effectively putting the observations in random order. Wanting a sample of size N, the user would finally create a subset of the sorted data set, the subset including only the first N of the randomly arranged observations.

Shown below is an example of selecting 50 observations randomly. Note that the use of the counting variable N is discussed on page 30 of the user's guide, and the UNIFORM function is described on page 246. We are supposing that the data collection BASE is stored on magnetic tape.

```
DATA POPULACE;
    INPUT DD=BASE ID 1-5 STATE $ 7-8 C 10-19 W $ 72-80;
    RANDOM=UNIFORM(43761);
PROC SORT;
    BY RANDOM;
DATA SAMPLE;
    SET POPULACE;
    N=N+1;
    IF N < = 50;
PROC PRINT;
    .
    .
    .
```

Taking a stratified random sample is more complicated. If we wish, say, to sample randomly 10% of the units <u>of each state</u> represented in BASE, we have to count how many observations are associated with each value of STATE. We show a sequence of SAS statements, annotated with COMMENT statements, that would accomplish the stratified random sampling. We recommend that a reader consult carefully the user's guide section on the MERGE statement, pages 47ff, if he desires a thorough understanding of the procedure.

```
         COMMENT
         WE BEGIN AS WE DID FOR SIMPLE RANDOM SAMPLING;
DATA POPULACE;
    INPUT DD=BASE ID 1-5 STATE $ 7-8 C 10-19 W $ 72-80;
    RANDOM=UNIFORM(34671);
         COMMENT
         AS WE SHALL TAKE 1/10 OF THE OBSERVATIONS OF EACH STATE,
         WE SORT "POPULACE" FIRST BY STATE.  THEN THE OBSERVATIONS
         WITHIN A STATE ARE PUT INTO RANDOM ORDER ACCORDING TO THE
         VALUES OF "RANDOM";
PROC SORT;
    BY STATE RANDOM;
         COMMENT
         NEXT, WE CREATE A PHONY DATA SET, CONTAINING NO OBSERVATIONS
         AT ALL, SO THAT WE CAN MERGE IT WITH "POPULACE" AND USE
         THE "LASTBY" FEATURES OF SAS'S MERGING OPERATION;
DATA PHONY;
    SET;
    STOP;
         COMMENT
         NOW, WE CREATE A VARIABLE "S_TOTAL" TO COUNT THE OBSERVATIONS
         WITHIN A STATE. WHEN THE "LASTBY" VALUE TELLS US THAT WE HAVE
         REACHED THE LAST OBSERVATION IN A STATE, WE RECORD ON THE DATA
         SET "COUNTS" THE VALUE OF "S_TOTAL", THE NUMBER OF OBSERVATIONS
         IN THAT STATE, AND RESET "S_TOTAL" TO ZERO.  THUS "S_TOTAL"
         CAN BE USED TO COUNT THE OBSERVATIONS IN THE SUCCEEDING STATE.
         THE DATA SET "COUNTS" WILL HAVE ONE OBSERVATION PER STATE;
DATA COUNTS;
    MERGE PHONY POPULACE;
       BY STATE;
    S_TOTAL=S_TOTAL+1;
    IF LASTBY THEN OUTPUT;
    IF LASTBY THEN S_TOTAL=0;
         COMMENT
         FINALLY, WE USE THE MERGE STATEMENT AGAIN TO PROCEED THROUGH "POPULACE" STATE
         BY STATE.  WE PUT INTO "SAMPLE" AT LEAST ONE AND NO MORE THAN 10 PERCENT OF
         THE OBSERVATIONS IN EACH STATE;
DATA SAMPLE;
    MERGE COUNTS POPULACE;
       BY STATE;
    N=N+1;
    IF N=1 OR N<=.1*S_TOTAL THEN OUTPUT;
    IF LASTBY THEN N=0;
PROC PRINT;
    .
    .
    .
```

If one wished further stratification of the sampling -- by values of STATE and C, for instance -- one need only substitute

```
    BY STATE C
```

everywhere

```
    BY STATE
```

appears above.

ANALYSIS OF VARIANCE IN SAS

The December, 1973, issue of the Journal of the American Statistical Association includes an article by Ivor Francis called "A Comparison of Several Analysis of Variance Programs." Mr. Francis describes a misuse of SAS's ANOVA procedure. We list below a few comments on the article.

1. The current documentation for ANOVA describes in detail the method used for calculating sums of squares (User's Guide, page 138 and pages 152-153). The method is considered appropriate for balanced factorial designs. The documentation does not claim it accurate for unbalanced factorial designs, such as the one treated in the article.* Mr. Francis, it should be noted, cites out-of-date SAS documentation in his criticism.

2. Mr. Francis claimed that no messages in the printout of the program suggested the cause of or solution to the problem noted. The ANOVA procedure, when confronted with an unbalanced design, always prints the message, "WARNING: THE ANALYSIS MAY BE INCORRECT DUE TO NO DATA FOR SOME CELLS. CHECK YOUR DEGREES OF FREEDOM AND CELL FREQUENCIES TO VERIFY THE CORRECTNESS OF THE ANALYSIS."

3. The table of contents of the User's Guide, as well as the REGR description, indicates that the REGR procedure is appropriate for analysis of variance, without regard to the balance of the design. Towards the end of his article, Mr. Francis acknowledges the correctness of the analysis produced by REGR. It should be noted that the claims made for the versatility of the BMD program apply equally well to REGR, and, as Mr. Francis wrote, REGR is "very simple to use in this fasion."

4. Mr. Francis ascribed the higher cost of running REGR to the presence of the other SAS procedures. The overhead is due rather to the sophisticated data management facilities of SAS. Also, many SAS jobs may be run in a memory region of 100K.

ANOVA is an efficient procedure for processing balanced designs and hence has its place in SAS. Like any computer program, it is open to abuse by careless users. Though we would have wished that Mr. Francis had emphasized much earlier in his article the appropriateness and virtues of SAS-REGR, we hope that his article will be of benefit in forestalling the misuse of SAS.

*For extremely large data sets of very slight imbalance, ANOVA's approximate results may be the only ones feasible to obtain.

We ask that the SAS installations copy and distribute this issue of SAS Communications. Users wishing to receive further issues are encouraged to send us their names and addresses.