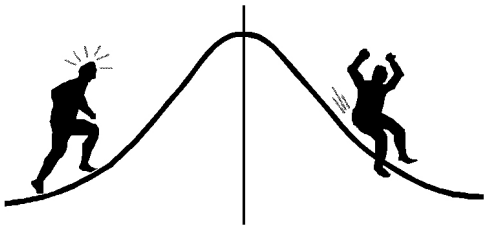

BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. # 20 DATE: July 4, 1989

SUBJECT: Rearranging Raw Data Files using SAS

Sometimes the data file you have in hand is not sorted conveniently for data analysis, or you may wish to select only some records for analysis. New raw data files can be easily and accurately created using SAS.

For instance, suppose the data requires two or more lines (or cards) for each record. Each line of a record is coded differently since different variables are on each line, and each type of line is identified by a cardtype variable. Lines of the same type have been coded together for greater ease and accuracy, but this results in a datafile where groups of cardtypes occur together. Data analysis requires that these cards be grouped contiguously so that they form individual records. Using an editing program (such as WYLBUR) to do this would be very time-consuming and liable to many errors.

For example, suppose that the identifying variables for each record are BLOCK, TREAT, and CARDTYPE, and that each line of data is 80 columns long. The following program solves this problem by using the `$Charw.` format:

```
DATA TEMP;
  INFILE 'A:\EXAMPLE.DAT';
  INPUT BLOCK 1-2 TREAT 5-6 CARDTYPE 7  @1 RECORD $CHAR80.;

PROC SORT DATA=TEMP;
  BY BLOCK TREAT CARDTYPE;

DATA TEMP;
  SET TEMP;
  FILE 'A:\SORTED.DAT';
  PUT @1 RECORD $CHAR80.;
RUN;
```

Knowledge of the actual coding for the different card types is not necessary with this program (except for knowing where the variables BLOCK, TREAT, and CARDTYPE are stored). SAS allows us to read the same characters twice so that we can specifically read in the values of BLOCK, TREAT, and CARDTYPE, and then re-read them as part of the RECORD variable (the `@1` tells SAS to go to column 1 and `$Char80.` tells SAS to read the next 80 columns).

The input dataset is called EXAMPLE.DAT on a diskette in A: drive, while the sorted data is stored as SORTED.DAT on the same diskette. The output dataset is sorted by ascending values of BLOCK. Within each BLOCK value, the TREAT values are in ascending order and within each BLOCK and TREAT values, CARDTYPE values are in ascending order.

The `$Charw.` format maintains all blanks, including leading blanks, so that the output record will look just like the input record. The `$Charw.` format is limited to 200 characters and is described in the SAS Language Guide for Personal Computers, Version 6 Edition on page 335.

Data can also be easily subset by including subsetting `IF` statements. For instance, if only data for BLOCKS 1 and 2 were to be analyzed then the following `IF` statement could be used

```
IF BLOCK = 1 OR BLOCK = 2;
```

in the first data step. More complicated subsetting statements could be used to tailor the output dataset as desired, even if sorting is not required.

Biometrics Information pamphlet #1 (Producing ASCII files with SAS) discusses another way to output ASCII files. Remember that the `PUT` statement requires that the number of digits to be put after the decimal place is indicated by a number after the column designations. If this number is missing, SAS assumes that it is zero (i.e. the numbers are truncated).

CONTACT: Wendy Bergerud
387-5676

PROBLEM FROM BI # 19

The pseudo F-value for moisture is calculated by:

$$F = \frac{54 + 1836}{108 + 162} = 7.0$$

with degrees of freedom calculated by:

$$df_{num} = \frac{\left[54 + 1836 \right]^2}{\left[\frac{(54)^2}{54} + \frac{(1836)^2}{2} \right]} = 2.119$$

and $df_{den} = 14.516$ as before. This F-value is now significant with $p = 0.0068$.