HOSPITAL SURVEY DATA MANAGEMENT USING SAS®

Phil Busby, UNC Department of Biostatistics

SENIC is an acronym for the Study on the Efficacy of Nosocomial (hospital-acquired) Infection Control, a large research project sponsored by the Center for Disease Control in Atlanta. This paper will present several features of SAS® we have found effective in processing data collected in the study.

The two basic goals of the SENIC Project are: 1) to provide statistical data that will describe the current extent and scope of infection surveillance and control program activity in U.S. hospitals; 2) to determine the degree to which infection surveillance and control programs have lowered nosocomial infection risks (i.e., incidence rates) over the past ten years.

The proposed method of data collection, called retrospective chart review, involves reading a patient's hospital chart after the discharge to determine if an infection occurred. In order to validate the chart review method, four pilot studies were conducted in which chart review results were compared with prospectively collected data.

SAS® may be used to merge in additional data such as that from the Hospital Interview Survey or the Regional Medical Program for cross-studies at the hospital level. The high level language features exemplified by the powerful file handling commands allow programmers to get out reports on subsets of a data base or crosstabulations on a few variables quickly. Usually the most time-consuming task is getting the data into a SAS® file. With several input formats to deal with, SAS saves a lot of conversion and reformatting work, as once the data is in a SAS® file, its location within the record is no longer a consideration, and a whole class of potential errors is eliminated.

The simplification of the data representation in SAS® results in faster implementation of systems not only due to reduced data conversion but also due to the reduction in the amount of communication needed among programmers. Use of SAS® on an interactive system such as TSO can double productivity in testing and editing with PROC EDITOR.

In the case of arrays within one observation, the repeated format list used in conjunction with the pointer specification facilitates some processing. In the following input statement example, 100 antibiotic dates are interwoven with 100 counts. The array of dates is read, then the pointer is reset to read the counts:

```
DATA SENIC.TYPSPEC:
INFILE H643DATA;
TITLE PSYS: PDC INPUT TO SAS;
INPUT
HOSPITAL $ 1-4
PATIENT $ 5-6
DCN $ 9-11
BEBEDM $ 12
BEBEDMNO $ 13
DPSEPNO $ 14-17
ADMISSION $ 18-19
ADMISSION $ 20-21
ADMITTY $ 22-23
DISCHRM $ 26-25
DISCHDD $ 26-27
DISCHTY $ 28-29
SERVICE $ 30
AGE $ 37-32
SEX $ 33
SITIN $ 38-39
OMSETH $ 40-41
TYPS $ 42
REDUCE $ 43
96 (ABDATT $ 10 - ABDATT99) (S$ 1, +2)
900 (ABCHT $ 10 - ABCHT99) (S$ 1, +5)
STDAT $ 89-89
STDAT $ 700-701
VCCODS $ 702
VCDATE $ 703-706
VCUT $ 707-708
VCDATE $ 709-710
VCWTR $ 711-714
VCWTR $ 715-716
(CCCMON $ 10 - CCCMON99) (S$ 1, +1)
A753 (SPECGR $ 10 - SPECGR10) (S$ 1, +3)
A779 (CATHID $ 10 - CATHID10) (S$ 1, +3)
A754 (PATHID $ 10 - PATHID10) (S$ 1, +3)
A777 (PATID $ 10 - PATID10) (S$ 2, +3)
A794 (PONAT1 $ 10 - PONAT10) (S$ 1, +3)
A750 (ADDUCT $ 10 - ADDUCT10) (S$ 1, +3)
A751 (PATHID $ 10 - PATHID10) (S$ 3, +3)
```

A useful enhancement would be an array operation procedure statement such as "LINK DOELCH FOR ITMME;ITMME;" where the subroutine DOELCH could have some notation like ITMME to represent the current index value.

The basis for the sample hospital selection was data from a preliminary screening questionnaire merged with the 1973 American Hospital Association tape. From this file certain hospitals were excluded for various reasons, and a stratified random sample was selected. These three tapes all had different record lengths and record counts. After input into the SAS® file all of the data was in one easily accessible place:
THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
SCHOOL OF PUBLIC HEALTH - DEPARTMENT OF BIOSTATISTICS
STUDY ON THE EFFICACY OF NOSOCOMIAL INFECTION CONTROL

PRELIMINARY SCREENING QUESTIONNAIRES
DESCRIPTIVE ANALYSIS DATA BASE CREATION

PHIL BUSBY
12/14/76

TITLE AHA PSQ ALL HOSPITALS SERVICE CODE 10, 44, 46-50, 56-59;
DATA WORK.EVERY;
FILE AHA PSQ:
INPUT
RANandom $ 1-8
PSQSTUdy $ 9-12
AMOUNT $ 13-19
TOTALS $ 29-36
TOTALM $ 39-46
QSTTEST $ 1065
QSTCLOSE $ 1066
QSTSCP $ 1067

LABEL RANandom = AHA RANDOM NUMBER HOSPITAL ID;
* FIRST DIGIT = '1' SPECIFYING AHA

LABEL PSQSTUdy = PSQ_study_number;
LABEL QHAVENTE = PSQ_HAVE_MEDICAL_EPHERIOLOGIST;
= L.1 QHAVENTE 0912 CAED=1,1L-6
* 1=Yes, 2=No, 3=No Answer

LABEL QMEERS = PSQ_ME HOURS пер. week;
* 1.2 QMEERS 0913-0914 CAED=1,CC=7-8
* 00-99, DECIMAL POINT ALLOWED, OR BLANK
* IF ANSWERS TO QUESTIONS 1.2 - 1.5 ALL BLANK,
* ANSWER 2= NO IS FORCED INTO QUESTION 1.1

LABEL QSTSCP = PSQ RESPONSE WITHOUT ISCP FLAG;
* 1=Yes, 2=No, BLANK=No
* TO BE COMPILED BY CDC

* NOTE ALL NUMERIC VARIABLES ARE STORED AS FLOATING-POINT NUMBERS;
* NOTE THE COMPARISON OPERATION 'IF NUM=1.3' DOES NOT YIELD A TRUE
RESULT IF THE VARIABLE NUM WAS GIVEN A LENGTH OF 4 AND ASSIGNED
THE VALUE 1.3 SINCE THE VALUE 1.3 IS A REPEATING DECIMAL FRACTION
AND WHEN THE COMPARISON IS MADE, THE VARIABLE NUM IS PADDDED WITH
ZEROS TO BRING IT UP TO THE SAME SIZE AS THE CONSTANT 1.3;

LENGTH CENSUS 8;
LENGTH TOTALIP 8;
LENGTH PAYIP 8;
LENGTH MOPATIP 8;
LENGTH TATIP 8;
LENGTH TEBEGIN 2;
LENGTH TPENDING 2;
LENGTH QSIGND 2;
LENGTH DEFAULT 4;

PROC SORT DATA=WORK.EVERY;
BY PSQSTUdy;
DATA WORK, UNIVER;  
* ___________________________________________________________________  
INFILE PSQADD;  
INPUT PSQSTUDY $ 9-12  
SUBSTUDY $ 1066-1069  
CONTROL $ 1070-1071  
PSCP $ 1072-1074  
SUBSTUDY $ 1075  
CONTROL $ 1076  
BEDSTAD $ 1077  
MEDSORT $ 1078  
* ___________________________________________________________________  
* THE FOLLOWING SEVEN VARIABLES WEREproduced BY THE PSQ  *  
* INDEXING AND STRATIFICATION PROGRAMS WRITTEN BY FRED D'IZIAZIO. *  
* INTERMEDIATE VARIABLES FROM WHICH RESULTS ARE DERIVED ARE: *  
* 1. MEDICAL EPIDEMIOLOGIST PERSONNEL ADJUSTMENT FACTOR; *  
* A. COLLECTING_ANALYZING_INTERPRETING_PER_BED_RATIO = *  
* (PSQ ME PERCENT HOURS CAI * PSQ ME HOURS PER WEEK); *  
* / AHA TOTAL BEDS *  
* B. TEACHING OTHER PER BED RATIO = *  
* (PSQ ME PERCENT HOURS TEACHING * PSQ ME HOURS OTHER) * PSQ ME HOURS PER WEEK); *  
* / AHA TOTAL BEDS *  
* C. FOR EACH RATIO ABOVE, DIVIDE DISTRIBUTION INTO 11 GROUPS *  
* ___________________________________________________________________  
LABEL BEDSTAD = PSQ HOSPITAL STRATUM NUMBER;  
* BEDSTAD = 1077  
* AHA TOTAL BEDS: 50-79  
* BEDRAK STRATUM: 1  
* 75-99  
* 100-149  
* 150-199  
* 200-299  
* 300-499  
* OVER 500  
* ___________________________________________________________________  
LABEL MEDSORT = PSQ MEDICAL SCHOOL SORT CODE;  
* IF TOTALS < 200, MEDSORT = '3'  
* ELSE MEDSORT = MEDSCHAP('11' OR '12')  
* ___________________________________________________________________  
PROC SORT DATA=WORk,UNIVER;  
BY PSQSTUDY;  
DATA WORK.PSQQNY;  
* ___________________________________________________________________  
UPDATE WORk.EVERY WORK,UNIVER;  
BY PSQSTUDY;  
DATA WORK.PSQSAMP;  
* ___________________________________________________________________  
INFILE SAMPLE;  
INPUT PSQSTUDY $ 9-12  
SAMPTRAT $ 1079-1084  
* ___________________________________________________________________  
MEMBER = PSQ SAMPLE STRATUM NUMBER;  
* ALL HOSPITALS SELECTED FROM THE UNIVERSE *  
* ARE SORTED BY PSQ SURVEILLANCE STRATUM NUMBER, *  
* PSQ SURVEILLANCE RELATIVE POS; *  
* PSQ CONTROL_RELATIVE POS; *  
* PSQ MEDICAL SCHOOL SORT_CODE; *  
* SEQUENTIAL NUMBERING IS DONE BY THE PROGRAM *  
* PSQSAMP BY STARTING WITH 1 AND INCREMENTING BY *  
* 1 EVERY TIME ONE OF THE VALUES OF THE ABOVE *  
* SORT FIELDS CHANGES FROM ONE HOSPITAL TO THE *  
* NEXT, THE HIGHEST VALUE OF THE SAMPLE STRATUM *  
* NUMBER GENERATED BY PSQSAMP ON 03/18/76 WAS 160. *  
* ___________________________________________________________________  
PROC SORT;  
BY PSQSTUDY;  
DATA SAS2.EVERY(READ=PAYA PROTECT=PAYA);  
UPDATE WORk.PSQQNY WORK.PSQSAMP;  
BY PSQSTUDY;  
* ___________________________________________________________________
The full specification of each variable in the above listing required extensive cross-referencing that was facilitated by the insertion of relevant information as comments in the PL/I record layout and under each SAS label statement. The PL/I variable name specifies the SAS variable label. Data entry keying corrections may be made by PROC EDITOR using the variable PSQSTUDY as the logical key.

The clarity and quality of the code can be improved further by structured walk-through of all code. Professional pride in the work is encouraged by printing the programmer name on all listings.

Crosstabulations on two sets of range values may be obtained by creating two new variables and assigning each one a descriptive character string corresponding to the value of the two variables for each observation. Since the FREQ procedure sorts the tables, spacing must be used within the descriptive character string to result in the desired printout sequence. This technique of dividing a task into small steps greatly augments programmer productivity.

If multiple SAS analyses are to be interspersed or arranged with other text, the output lines may be routed to a tape for subsequent input into a text editing system using PROC PRINTTO as illustrated in the following example:

```plaintext
I/SAS2 DSN=UNC.E.P2306.SENIC.SAS.DEVFERV.UNITYR=DISK.
// DISPIDLD.VOL=SER=USSASS
//F20DDDD DSN=UNC.E.P2306.SENIC.SAS.PSQ13.UNIT=TAPE.
// DISPIDLD.KEEPAING=IN,LABE=1153,IVERY=UT829.
//SYSS DD *

*************** AHA PSQ DESCRIPTIVE ANALYSIS 1.3 ***************;
* THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL *
* SCHOOL OF PUBLIC HEALTH - DEPARTMENT OF BIOSTATISTICS *
* STUDY ON THE EFFICACY OF INFECTION CONTROL *
* PRELIMINARY SCREENING QUESTIONNAIRE *
* DESCRIPTIVE ANALYSIS OF SECTION I, QUESTION 3 *
* LINDA HODG *
* 11/30/76 *
*

THE FOLLOWING PROCEDURE WRITES THE OUTPUT TO A TAPE:
PROC PRINTTO NEW UNIT=20;

TITLE1: GALLA13;
TITLE2: AHA PSQ ALL HOSPITALS SERVICE CORE 10.44.66-50.56-59;
TITLE3: CROSSTABULATIONS ON SECTION 1, QUESTION 3;
TITLE4: TIME PERCENTAGE OF INFECTION CONTROL ACTIVITIES;
TITLE5: FOR EACH MEDICAL EPIDEMIOLOGIST;

DATA WORK.QMEPH;
SET SAS2.EVERY(READ=PAY);

* THE FOLLOWING STATEMENT OMITS ALL OBSERVATIONS FOR HOSPITALS WITHOUT A MEDICAL EPIDEMIOLOGIST;
IF QMAYEHE = '1';
```
The following three statements correct any percentage of hours entered as 99 with the other two percentages being, in which case it is assumed that the respondent would have entered 100 percent but could not because there were only two digits allowed on the form:

```
IF QMPCICA >= 99
  & QMPPBTA = 0
  & QMPHOTH = 0
THEN QMPCICA = 100;
IF QMPPBTA >= 99
  & QMPCICA = 0
  & QMPHOTH = 0
THEN QMPPBTA = 100;
IF QMPHOTH >= 99
  & QMPCICA = 0
  & QMPPBTA = 0
THEN QMPHOTH = 100;
```

The following statements assign the descriptive character string value corresponding to the range of the medical epidemiologist's percentage of time spent on collecting, analyzing, and interpreting infection control data to the SAS variables called 'collect' created here to be cross-tabulated with the traching percentage range values by the FREQ procedure below:

```
IF QMPCICA = 0
THEN COLLECT = '0';
IF QMPCICA > 0 & QMPCICA < 10 THEN COLLECT = '1 TO 9';
IF QMPCICA >= 10 & QMPCICA < 25 THEN COLLECT = '10 TO 24';
IF QMPCICA >= 25 & QMPCICA < 50 THEN COLLECT = '25 TO 49';
IF QMPCICA >= 50 & QMPCICA < 75 THEN COLLECT = '50 TO 74';
IF QMPCICA >= 75 & QMPCICA < 100 THEN COLLECT = '75 TO 99';
IF QMPCICA >= 100
THEN COLLECT = '100';
```

The following statements assign the descriptive character string value corresponding to the range of the infection control to the SAS variable called 'teach' created here to be cross-tabulated with the collecting-analyzing-interpreting range values by the FREQ procedure below:

```
IF QMPPBTA = 0
THEN TEACH = '0';
IF QMPPBTA > 0 & QMPPBTA < 10 THEN TEACH = '1 TO 9';
IF QMPPBTA >= 10 & QMPPBTA < 25 THEN TEACH = '10 TO 24';
IF QMPPBTA >= 25 & QMPPBTA < 50 THEN TEACH = '25 TO 49';
IF QMPPBTA >= 50 & QMPPBTA < 75 THEN TEACH = '50 TO 74';
IF QMPPBTA >= 75 & QMPPBTA < 100 THEN TEACH = '75 TO 99';
IF QMPPBTA >= 100
THEN TEACH = '100';
```

The following statement keeps variables collect & teach. The other variables are ignored:

```
KEEP COLLECT TEACH;
```

```
PROC FREQ;
TABLES COLLECT TEACH;
```

The program statements above generate tables for the entire data base. Subsets are obtained by the SAS IF statement. Then TITLE and SET statements may be changed for each run using TSO editing commands or SAS MACRO statements. The first title line contains the key for each set of results retrieved by the text editing system.
TABLE OF COLLECT BY TEACH

<table>
<thead>
<tr>
<th>COLLECT</th>
<th>TEACH</th>
</tr>
</thead>
<tbody>
<tr>
<td>FREQUENCY</td>
<td>PERCENT</td>
</tr>
<tr>
<td>COL PCT</td>
<td>D</td>
</tr>
<tr>
<td>0</td>
<td>---</td>
</tr>
<tr>
<td>1 TO 9</td>
<td>---</td>
</tr>
<tr>
<td>10 TO 24</td>
<td>---</td>
</tr>
<tr>
<td>25 TO 49</td>
<td>---</td>
</tr>
<tr>
<td>50 TO 74</td>
<td>---</td>
</tr>
<tr>
<td>75 TO 89</td>
<td>---</td>
</tr>
<tr>
<td>90 TO 99</td>
<td>---</td>
</tr>
<tr>
<td>TOTAL</td>
<td>---</td>
</tr>
</tbody>
</table>

Bibliography


McClish, Donna, and Quade, Dan, "On the Accuracy of Retrospective Chart Review," Reprois 36, SENC Project technical papers, UNC Dept. of Biostatistics, 1977