PROC SUMMARY as the Basis for Efficient Analyses of Large Files

Juliana M. Ma, UNC Highway Safety Research Center
Carol Leininger, Research Triangle Institute

Introduction

A strategy for efficient analyses of large files can be based on data generated by PROC SUMMARY. When a study involves only a few categorical variables, then several analyses can be based on one set of stored summary counts. The strategy description includes planning recommendations to increase the potential usefulness of summary data.

An outline of the strategy, its advantages and disadvantages, follows a brief description of the data files used. Some familiarity with PROC SUMMARY is assumed. Examples of actual applications are provided.

Background: File Description

The files used to develop the strategy are the North Carolina Accident files from 1973 to the present, the major database at the UNC Highway Safety Research Center. Each accident year has about 300,000 records. Since 1973 there have been two different data formats. The format from 1973 to 1978 includes 110 variables; beginning in 1979 there are 150 variables. Many variables are included in both formats.

Analyses of our accident data often span several years, and frequently involve less than ten variables. One to six years may be processed as one file depending on the complexity of the processing.

Applications

The strategy is useful for analyses that only involve a few categorical variables. This is often the case for our accident data. Some frequently used variables are Occupant Injury (5 non-missing levels), Seat Belt Usage (2-4 non-missing levels), Age (usually less than 10 categories), Vehicle Type (2-5 categories) and Accident Severity (5 levels).

Another application involves indicator (0,1) variables. Rare events such as vehicle roll-overs can be counted more efficiently using an indicator variable instead of a classification variable. The incidence of a rare event can be determined for sub-groups defined by categorical variables. For example, post-crash fires can be studied for different types of cars as defined by Manufacturer, Car Type and Model Year.

The final analysis products are tables of counts and appropriate percentages. Results can be obtained from PROC FREQ or PROC TABULATE for standard cross-tabulations. Customized SAS® programs may be required to generate special tables when indicator variables are used.

Strategy

Advance planning is the key to optimum use of summary counts. The actual process of using PROC SUMMARY to generate counts for later analyses is straightforward. The most important decisions involve the choice of variables and the number of categories to allow for each variable.

As a first step, study the analysis request. Make a list of all the variables required for tabulations. These will be included in the CLASS statement. The list should not include variables only used for record selection. For example, Vehicle Type would not be included if it were only used to select passenger car records.

After specifying the required variables, the optimum number of levels of each variable should be chosen. Ideally, all possible levels of all variables should be included to provide maximum flexibility in later analyses. When this is not possible, the number of levels combined should be kept at a minimum. For instance, if Age is a primary study variable and the research emphasis is on children, then ages less then ten could be left unchanged and only older ages grouped into categories.

Regardless of the number of variables, the potential maximum number of cells should be less than 20,000. PROC SUMMARY can handle up to 32,000 cells, but our experience has shown that summary datasets of more then 15,000 cells are not worth the trouble. The absolute maximum number of cells can be found by multiplying the number of levels for each categorical variable. For example, Injury (5 levels), Belt Usage (2 levels) and Sex (2 levels) could produce at most twenty (5x2x2) cells. The actual number of cells

usually will be smaller than the maximum number, especially if some levels are rare (fatal injury for example).

Indicator variables can be created for counting rare events, and put in the VARIABLES statement. The SUM option of the OUTPUT statement will generate summary counts of the indicator variables for each sub-group defined by the categorical variables included in the CLASS statement.

Summary counts should be stored in a permanent SAS dataset when they are created. The NWAY option should always be included in the PROC SUMMARY statement to suppress marginal counts. The MISSING option should be included if any value may be missing since SUMMARY deletes the entire record if any variable has a missing value. List all the categorical variables in the CLASS statement. If indicator variables are listed in the VARIABLE statement, the OUTPUT statement must include the SUM option. Allow ample memory, at least 600K, for processing. The _TYPE_ variable may be dropped from the output dataset to save storage space.

The last, and least expensive, step is to use stored summary counts to produce a variety of tables. The _FREQ_ variable is used as the "weighting" variable for PROC FREQ or PROC TABULATE. Customized programs may be necessary to calculate percentages when indicator variable counts are computed.

Advantages

This strategy has several advantages. Analysis of summary data can reduce processing costs. PROC SUMMARY is an efficient way to compute compute summary counts. Sharlin (SUGI '83) found that "PROC SUMMARY should be the 'default' choice for computing summary statistics" (p. 919) after comparative testing. Even small files may be processed more efficiently using PROC SUMMARY (see Example 3).

With proper planning, one pass of a large file can provide summary data useful for several analyses. Since any combination of variables included in the CLASS statement can be analyzed, a variety of cross-tabulations can be produced. Subfiles may be studied by using CLASS variables to create subsets. For instance, including Region of Impact would allow analysis of rear-struck vehicles only.

Specifying the NWAY option reduces the number of summary records produced, with no loss of information. Marginal counts can always to generated by using the summary data as input to an appropriate SAS Procedure.

Extremely large files can be processed in sections to create summary datasets, which can be combined for the final tables. We usually do not process more than six years of accident data at once, but often want tables that include information for ten or more years. Dividing the processing is also useful when information comes from files with different formats.

Careful selection of variables and category collapsing may create summary data useful for low-cost, previously unplanned analyses. If the values for a variable are not collapsed then tables with a different grouping can be produced without processing the large file again. If the total number of requested variables is small, potentially related variables could be included for little extra cost.

Using summary data can make program testing easier, especially when the exact format of the final tables is important. Tables based on test data generated by using only the OBS option often have empty or non-existent cells for rarer combinations. A test summary dataset can be generated from 10,000 to 50,000 records at relatively low cost to produce better test data. Alternatively, table formatting can be based on actual summary data.

Generating tables using summary data is generally less expensive than processing data directly; this is most obvious for large files. Once the summary data is stored, any table based on the variables included in the CLASS statement can be easily created.

Disadvantages

The most obvious difficulty with this strategy is the limitation on the number of variables (and number of levels). Since all variables must be included in the CLASS statement, usually the limit is five or six depending on the number of levels. A BY statement may be used for key variables if the file is sorted appropriately.

The summary data may not always be useful for later unanticipated analyses. Collapsed categories cannot be separated. If a variable is omitted, then the entire large file must be reprocessed unless an appropriate subfile was created. Of course, omitting variables can be a problem with other methods as well.

The MISSING option must be included if any values are missing. Omitting this option when missing values are present may produce cross-tabulation couts that are misleadingly small. PROC SUMMARY deletes any record with a missing value. A valid record will be be included in a two-way table based on summary data if any variable in the CLASS list is missing in the record and the MISSING option is omitted.

Example 1

A study of children in car accidents, ages less than four, with regard to the NC Child Restraint Law begun in July 1982. The data files are extracts of three years of NC accident data, 1980-1982 (occupants less than four years old, 24,196 records).

Class variables	Number of levels
PERIOD	4
SEAT	6
TADSEV	8
AGE	4
BELTUSE	6
INJURY	6

Maximum number of cells: 27,648
Actual number of cells: 3,310

The original request only included four variables for all occupants under age four: Time Period (80, 81, 82 before, 82 after), Age (0-1, 2-3), Belt Use (No restraint, Restraint used) and Injury (5 levels, KABCO). Age and Belt Use were not collapsed in the summary data although the requested tables only required two levels for both variables. This allows later analyses to have different groupings for these variables. Seat Position and TAD Severity (a measure of vehicle damage) were included since they were used in previous analyses.

The actual number of cells was much less than the maximum possible number since the more severe levels of TAD Severity and Injury are rare. Less than 10,000 cells were expected based on previous experience with the variables.

Final tables were produced by PROC FREQ and PROC TABULATE. PERIOD, SEAT and AGE could be used as selection variables.

Example 2

A study of Make/Model differences in cars for rare events (Rollover, Post-crash fire). The data are extracts of compatible vehicle variables from multi-state data (NC, MD, NY, SC), ten years from NC, four to five years from the other states (small cars, model years 74-78, 372,158 records).

BY variable: STATE

Class variables	Number of levels
MAKE	20
VINYR	5
BODYTYPE	5
CARLINE	58
IMPACT	7

Maximum number of cells:
10,150 per state for 4 states
Actual number of cells: 8,370

MAKE and CARLINE are redundant because CARLINE is defined uniquely within each MAKE. Thus the maximum number of cells is 5x5x58x7, not 20x5x5x58x7. The actual number of cells was reduced further because IMPACT included relatively rare levels and some car models do not exist in all VINYRs (model years) or BODYTYPEs.

Indicator variables ROLLOVER and FIRE were created (1 if event occurred, 0 otherwise). Different variables were used in each state to define comparable Rollover and Post-crash Fire indicators.

A custom program that computes percentages for the indicator variables was used to produce tables in addition to PROC FREQ. CARLINE and BODYTYPE were used to select car models included in the final tables.

Example 1

```
PROC SUMMARY DATA=CHILD  NWAY MISSING;
    CLASS  PERIOD SEAT TADSEV AGE BELTUSE INJURY;
    OUTPUT OUT=ddname.STAT(LABEL='NC 80-82 AGE<4');
PROC TABULATE DATA=ddname.STAT  F=8.;
    CLASS  AGE PERIOD BELTUSE INJURY;
    FREQ   FREQ_;
    FORMAT AGE AGECAT.  BELT BELTCAT.;
    TABLES BELT * INJ,  PERIOD * (AGE ALL='0-3 YEARS');
```

Example 2

```
PROC SUMMARY DATA=CARS  NWAY;
  BY STATE;  /* CARS created in order by STATE */
    CLASS  MAKE VINYR BODYTYPE CARLINE IMPACT;
    VAR  ROLLOVER FIRE;
    OUTPUT OUT=ddname.STAT(DROP=_TYPE_)
           SUM= ;
```

Example 3

Run	SAS Log Time				MVS System Time				Cost
	Total	DATA	SUMMARY	TABULATE	Total	CPU	I/O	Other	
Standard	3.63	0.52		3.11	24.0	12.4	8.5	3.4	$3.44
Comparison	2.71	0.53	0.68	1.50	22.1	10.1	8.5	3.5	$3.11

Example 3

A cost comparison using a "small" file with 1648 observations, 21 variables and a table with five categorical variables. The standard program included a DATA step and PROC TABULATE. The comparison program had the same DATA step, PROC SUMMARY and PROC TABULATE using the summary data.

Conclusion

This strategy has proven useful for many analyses of our accident files. Basically, this is an expansion of the advice given in the SAS Applications Guide for processing large datasets. We hope some variation of the ideas presented here can be applied to other files.

Note 1

The programs were run during standard weekdays hours at the Triangle Universities Computation Center, Research Triangle Park, NC.

References

Council, Kathryn A., editor, SAS Applications Guide, 1980 Edition. SAS Institute Inc., Cary, NC, 1980. Chapter 10: Processing Large Data Sets with SAS, pp. 149-157

Sharlin, Joshua, "Data Reduction and Summarization." Proceedings of the Eighth Annual SUGI Conference, SAS Institute Inc., Cary, NC, 1983, pp. 912-919.

Hamilton, Elizabeth, Single Variable Tabulations for 1975-1978 North Carolina Accidents, UNC Highway Safety Research Center, Chapel Hill, NC, 1979.

Hamilton, Elizabeth, Single Variable Tabulations for 1979-1981 North Carolina Accidents, UNC Highway Safety Research Center, Chapel Hill, NC, 1982.

For more information, contact:

Juliana M. Ma
UNC Highway Safety Research Center
CTP-197A
Chapel Hill, NC 27514
919-962-2202