Paper 041-29

Proc FREQ –What's it really good for?

Theresa Gordon, US Census Bureau, Suitland, MD
Monique Eleby, US Census Bureau, Suitland, MD

Abstract

Proc FREQ is a procedure that is used to give descriptive statistics about a particular data set. Proc FREQ is used to create frequency and cross-tabulation tables. It enables analysis at various levels. Associations between variables and responses can be tested and computed. Proc FREQ is widely used to analyze healthcare datasets and demographic databanks that contain specific information about individuals and their activities. Any dataset that requires analysis on an individual or multi-variable level can be manipulated with this procedure. Common user pitfalls and frequent misuses of the proc FREQ procedure will be demonstrated and discussed. In summary, we will re-address the basic points of proc FREQ and reiterate the things to avoid when using proc FREQ, and address any questions that may be raised.

Introduction

Proc FREQ can be used to create one-way, two-way or n-way tables. These tables are useful in analyzing the various values of a particular variable and their contributions to the overall dataset as a whole or subset therein. A two-way cross-tabulation table provides a summarization for two variables in a dataset. It provides an analysis tool for mutually exclusivity and a percentage break down of the various combinations of responses. An n-way cross-tabulation table provides three or more multi-way tables that allow an individual to analyze variable combinations at a greater depth. For the sake of simplicity, we will limit our presentation to just one-way and two-way tables.

Overview

Basic Syntax for a two-way table:

```
Proc FREQ <options>;
        BY (pre sorted variable 1) (pre sorted variable 2);
        TABLES variable 1 * variable 2;
        WEIGHT;
Run;
```

A two-way table is useful where relational statistics for the two named variables are desired so that the intersection, exclusion, and combinations of data can be analyzed.

Figure 1.

		Do you like hot dogs?		
		Yes	No	
Do you like Pizza?	Yes	27	36	63
	No	5	32	37
		32	68	100 Total

Figure 1 is an example of two-way frequency table that is based on a survey of 100 elementary school kids who were asked Question #1: whether they like pizza, and Question #2: whether they like hot dogs. The responses to Question #1 are listed down the vertical side of table and the responses to Question #2 are listed along the horizontal side of the table. The number of students who liked both pizza and hot dogs was 27. The number of students who like pizza but not hot dogs is 36. The number of students who like hot dogs and not pizza was 5 and those that didn’t like either numbered 32. The total of both rows and the total of both columns should equal the total number of cases at hand.

Things to Avoid/Pitfalls

Proc FREQ’s Default Truncation

If a character variable value is longer than 16 characters, Proc FREQ will automatically truncate the variable to 16 characters. This can provide the user with incorrect results, because after truncation, like values are combined.

For example:

Variable 1: Today I am going to the store
Variable 2: Today I am going to the basketball game
Variable 3: Today I am going past the dry cleaners

Each one of these variables will be truncated to the value “Today I am going” and no difference between them will be recognized.

In order to avoid such an error, Use Proc SUMMARY with the NWAY option and print your results so that all variables will be displayed in their full context.

Formatting Variables with Missing Values

Proc FREQ is a procedure that formats variables before it determines the number of missing values. This could cause missing and some non-missing values to be formatted the same way and erroneously be considered equivalent resulting in a loss of data accuracy and integrity.

In order to avoid this error, when creating formats for variables, include a separate range or designation for missing values in the Proc FORMAT value assignment. This will allow the missing values to be recognized and coded during the formats therefore resulting in a proper assignment of formatted values.

Proc FORMAT example:

```
Proc FORMAT;
        Value Correct     .  = “MISSING”
                          P  = “PIZZA”
                          H  = “HOTDOG”
                          B  = “BURGER”;
Run;
```

Memory Limitations

When processing a FREQ procedure it is possible to run out of memory due to the fact that this procedure stores all of the value combinations in memory during processing. Processing multi-way tables and/or multi-level variables can prove to be space intensive and consume all available memory.

In order to prevent this error, try using Proc SORT to sort the data by as many variables as you can and also include those sorted variables in a BY statement in your Proc FREQ procedure. A possible solution to this issue could also be an upgrade to a higher level of memory.

Table Level Limitations

Although Proc FREQ can be manipulated and configured to process numerous tables of requests containing numerous variables, the procedure does have a limit to its level processing ability. Proc FREQ allows a maximum of 32,767 levels for any one individual variable.

In order to process more than 32,767 levels of variables, opt to use the Proc SUMMARY procedure to obtain the desired statistics. A format statement can also be used to reduce the number of levels being processed by the FREQ procedure.

Conclusion

Proc FREQ is an extremely useful statistical tool in the SAS language. This procedure outputs Frequencies, cross-tabulation tables, various measures of variable associations across a data set, and stratified analysis. Although it is extremely applicable in the analysis of data, the user must have a proper understanding of the limitations and precautions that accompany this procedure prior to utilizing it. A thorough understanding will ensure that the integrity and accuracy of the data is not compromised during the data analysis phase.

References

Cody, Ronald P., and Pass Raymond, *SAS® Programming by Example,* Cary, NC: SAS Institute Inc., 1995.

Mason, Phil, *In the Know…SAS® Tips and Techniques from Around the Globe*, Cary, NC: SAS Institute Inc., 1996.

SAS Institute Inc., *SAS® Procedures Guide, Version 8, Volume 1,* Cary, NC: SAS Institute Inc., 1999.

Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

Theresa L. Gordon
U.S. Census Bureau
4700 Silver Hill Road
Suitland, MD 20746
Work Phone: (301) 763-6961
Fax: (301) 457-3932
Email: Theresa.L.Gordon@census.gov

Monique E. Eleby
U.S. Census Bureau
4700 Silver Hill Road
Suitland, MD 20746
Work Phone: (301) 763-6962
Fax: (301) 457-3932
Email: Monique.E.Eleby@census.gov