

USING SAS TO ANALYZE SURVEY RESULTS

Stacey G. Bricka

Center for Urban Transportation Research (USF)

ABSTRACT

Surveys are a popular method of collecting perceptions, attitudes, and opinions on a variety of topics. The analysis of survey results can be accomplished through the use of SAS. The main techniques for producing descriptive statistics are through the use of frequencies and means. These procedures can be performed on the entire sample, or on subsets if desired. Tests for statistical significance can be performed on the data, as can more detailed statistical analyses, such as factor analysis. Using SAS/Graph, the results of such analyses can be readily displayed.

The intent of this paper is to present the basics of survey analysis using SAS. Sample programs, where appropriate, will be presented using PC-SAS programming code. Specifically, data entry, validation, analysis, and presentation will be discussed, in order to assist beginning users of SAS discover the benefits of this type of application. This discussion is not inclusive of all methods available for survey analysis. For more detailed descriptions of the procedures discussed, consult the SAS Procedures Guide.

DATA ENTRY

Data entry can be accomplished through the use of SAS/FSP, Lotus 1-2-3 (Lotus), or other similar software packages. SAS/FSP, as described in the SAS Language Guide, is an interactive, menu-driven facility for data entry. Through the use of SAS/FSP,

it is possible to restrict the data entry process, by defining the data entry fields as character or numeric and by setting a maximum field length. This will assist in data validation.

If SAS/FSP is not available, it is possible to enter the data using another software program and convert the data into a SAS dataset. One such software program is Lotus. If using Lotus, the spreadsheet should be constructed such that the first row contains all the variable names, with one name per column. The observations are entered in the following rows, one record per row. After the data is entered, it is necessary to translate the data into a dBase or ASCII file, in order to convert the data into a SAS dataset. Conversion of dBase or ASCII files into SAS datasets can be accomplished by using PROC DBF for dBase files or PROC DIF for ASCII files. These procedures are discussed in the SAS Procedures Guide.

DATA VALIDATION

Once the SAS dataset has been created, the data can be checked for accuracy through the use of the frequency procedure (PROC FREQ). This procedure will show the distribution of values for each variable. A sample SAS program to produce frequencies for the dataset "SURVEY" is:

```
DATA SAMPLE;  
SET SURVEY;  
RUN;  
PROC FREQ;  
RUN;
```

The values shown should be within the range of appropriate responses (e.g., if the responses are between 1 and 6, there should not be any 0's, 7's, or character values). If there are inappropriate values, it is possible to trace these to the original survey by printing the dataset (through the use of a PROC PRINT). After the correct response is determined, the dataset must be corrected. This may be done by going into the data entry software and correcting the entry, or it is possible to use a simple SAS program. For example, suppose the variable "Q1A" should have values of "1" through "6", but contains one "7". By checking the original survey (identified by ID=2598), it is found that the actual response should be "6". The SAS program to correct the dataset "SURVEY" would be:

```
DATA DATA.SURVEY;
SET SURVEY;
IF Q1A=7 AND ID=2598 THEN Q1A=6;
RUN;
```

If any mathematical calculations are to be performed on the data, the variables must be formatted as numeric. The procedure used to print the contents of the dataset (PROC CONTENTS) will list the dataset variables, the length of the field, and if the field is defined as character or numeric.

ANALYSIS

Once the data have been validated for accuracy, it is possible to begin the analysis. This can be as simple or complex as desired. For the purposes of this paper, two simple techniques will be presented, frequencies and means. In addition, one technique for determining statistical significance will be presented.

As discussed above, using the frequency

procedure (PROC FREQ) will provide a distribution of responses for each variable in the dataset. A more in-depth look into the responses can be accomplished by performing a cross-tabulation or "cross-tab". For example, to determine how gender compares with responses to a particular question, a cross-tab could be performed on the variables GENDER and QUESTION. A sample program might be:

```
DATA CROSSTAB;
SET SURVEY;
RUN;
PROC FREQ;
TABLES GENDER*QUESTION;
RUN;
```

The output from this program would be a table with the responses to QUESTION across the top of the table, and the responses to GENDER down the side. Cross-tabs allow a more focused examination of the data, and are useful in interpreting the distribution produced by PROC FREQ.

A second analysis technique is the means procedure (PROC MEANS). In order to perform this procedure, the data must be numeric. The program for calculating the mean responses for each survey question is:

```
DATA MEANS;
SET SURVEY;
RUN;
PROC MEANS;
RUN;
```

The output for this program will show, for each variable, the number of observations in the dataset, the variables in the dataset, the number of responses corresponding to each variable, the minimum and maximum values, the mean, and the standard deviation about the mean. It is possible to tailor the output. For example, if the desired output would be to show the number of responses, the mean, and the

standard error about the mean, then the program would be:

```
DATA MEANS;  
SET SURVEY;  
RUN;  
PROC MEANS N MEAN STDERR;  
RUN;
```

(Consult the procedures manual for all available options in the means procedure.)

As with PROC FREQ, it is also possible to evaluate in more detail the results obtained with PROC MEANS. This can be accomplished through the use of a "CLASS" (or classification) statement in the program. The CLASS statement follows the procedure line in the above program. By specifying a classification variable, such as GENDER, the resulting output contains the means of all responses which had a GENDER response of female, followed by those with GENDER response of male. This is similar to using a TABLES statement in the frequency procedure.

In order to determine statistical significance of the responses, the MEANS and TTEST procedures can be used. If the objective is to determine if a particular response is statistically different from zero, the MEANS procedure can be used. Simply include "T" as an option in the procedure line of the program. If the goal is to determine if responses by the men are statistically different from those of the women, the TTEST procedure is useful. This procedure tests to see if the means of the responses by each group are equal. A sample program would be:

```
DATA TTEST;  
SET SURVEY;  
RUN;  
PROC TTEST;  
CLASS GENDER;  
RUN;
```

The resulting output would enable the researcher to determine if responses by the male respondents are statistically different from those of the female respondents.

The above discussion has suggested simple methods for evaluating questionnaire responses. Once the data have been analyzed, it is often necessary to convey the results to others.

PRESENTATION OF RESULTS

The survey results can be graphically presented through the use of a standard graphics software, such as Harvard Graphics or Charisma. In addition, SAS/GRAPH has excellent capabilities in this area.

Through the use of any of these packages, it is possible to convey the results to others. Standard methods for presenting the results include bar charts, line graphs, and pie charts. The decision of which method to use will rest with the type of data being presented. It may be useful to display the results using several methods, in order to determine which best conveys the results.

If the purpose of the graphic is to display the distribution of responses to a particular question, a bar chart could be used. The program to produce a bar chart displaying the frequency of responses to Question 1a (or variable Q1A) would be:

```
DATA BARCHART;  
SET SURVEY;  
RUN;  
PROC GCHART;  
VBAR Q1A/DISCRETE;  
RUN;
```

Consult the SAS/GRAPH Guide for information on how to present data using other methods.

CONCLUSION

This paper has presented techniques for analyzing survey data using SAS. It is intended to serve as an introduction, and has relied greatly on the information presented in the SAS Procedures Guide and the SAS Language Guide.