Paper RIV15

SAS® Macros to Produce Publication-ready Tables from SAS Survey Procedures

Emma L. Frazier, Centers for Disease Control, Atlanta, Georgia
Shuyan Zhang, ICF International, Atlanta, Georgia
Ping Huang, ICF International, Atlanta, Georgia

ABSTRACT

To analyze complex survey data, analysts must have the knowledge of weights and design variables required to complete the analysis. SAS® Survey procedures are used for the analysis of this type of data, but the output can be challenging for the production of quality tables. We developed SAS code that uses the features of the SAS® ODS and Proc Report to generate publication-ready documents from SAS Survey procedures to complement data analysis for end users. We present relatively straightforward SAS code that will generate rich text format tables for cross tabulations with statistics such as weighted percentages, 95% confidence intervals, coefficient of variation, and Rao-Scott Chi-Square p-values.

Some advanced options and techniques in PROC REPORT are presented to demonstrate the flexibility of customizing the output style. The program has a simple interface with the capability to create complex table formats. These procedures are valuable for researchers who need to produce tables for analysis and can be easily modified for various tables.

INTRODUCTION

Data from surveillance systems often involve a large number of variables and are often collected using a complex survey design. To analyze data that were collected using complex survey methods, analysts must employ statistical methods that incorporate weights and design variables, e.g., strata and cluster variables. SAS has developed SAS Survey procedures to permit the analysis of data collected using complex sampling methods. While these procedures are effective in producing weighted analyses that adjust for the sampling schemes, the outputs from these procedures often are voluminous and produce a large number of pages of SAS output.

Data analysts and researchers are often challenged to sort through pages of output to produce tables. It is often inefficient to review pages of rows of output and manually re-enter data into tables to produce the desired formatted output. To reduce the burden of transferring data from numerous output pages into tables, a PROC REPORT procedure, along with Output Delivery System (ODS) statements, and macros were developed to produce publication-ready tables using the outputs from SAS Survey Procedures. This process was developed to help streamline the table production process and develop clear, well-defined publication-ready tables, including all required, relevant statistics.

This paper will provide a brief overview of the data used in our analyses and the SAS® Survey procedure, and will present the relevant parts of the SAS code and macro to produce publication-ready tables.

DATA SOURCE

The example analysis in this paper uses data from the Medical Monitoring Project (MMP), a supplemental surveillance project designed to produce nationally representative estimates of behavioral and clinical characteristics of HIV-infected adults receiving outpatient medical care in the United States (Blair, McNaghten et al. 2011; Frankel, McNaghten et al. 2012). MMP is a complex, three-stage probability-proportional-to-size (PPS), cross-sectional survey. For the 2009 data collection cycle, first U.S. states and territories were sampled, then facilities providing HIV care, and finally persons aged 18 years or older receiving at least one medical care visit in participating facilities between January and April 2009. Data were weighted based on known probabilities of selection at state or territory, facility, and patient levels; weights were also adjusted for non-response using predictors of patient-level response.

Strata and clusters were employed to reflect the study design and reduce bias in the variance estimates. Data were weighted to produce national estimates of HIV-infected persons in care.

MMP provides a rich data source to increase the information about the health and health care of adults who are infected with HIV. Numerous key data elements exist, including data related to socio-demographic factors (e.g., race/ethnicity, education, income, gender), to access to care (e.g., regular place of care, date of last HIV care), to unmet needs (e.g., housing, child care), and to behavioral factors (e.g., smoking, drinking, use of illicit drugs).

Description of the Problem

Given the wealth of MMP data elements and information, researchers are frequently faced with producing publication-ready tables of various formats to disseminate data from the MMP dataset. One critical job of the data analyst is to facilitate researchers to produce these publication-ready tables in a timely manner. Prior to the availability of this code, researchers manipulated the SAS output in Microsoft Word or Excel and transcribed the output by hand to produce the tables with the special formatting requirements. This code has reduced the burden of producing tables that are easily understood by the researcher and lay persons, while preparing tables that are ready for publication.

For purposes of this paper, we will focus on a series of bivariate analyses and provide selected statistics to investigate the relationship between an outcome and the independent variables. An example of an abbreviated table shell for a publication-ready table is provided below.

Table Shell for Publication-ready Table
Table 1. Prevalence of smoking among HIV-infected persons in care, by selected characteristics - Medical Monitoring Project, United States, 2009

	Total		Smoker		Non-smoker		
	N	% (95% CI)	N	% (95% CI)	N	% (95% CI)	p-value
Demographic factors							
Age							
18-29							
30-39							
40-49							
≥50							
Gender							
Male							
Female							
Race/Ethnicity							
White, non-Hispanic							
Black, non-Hispanic							
Hispanic or Latino							
Other							
Education							
<High School							
High School diploma							
>High School							

The SAS SurveyFREQ will permit us to generate selected statistics such as weighted percentages, 95% confidence intervals, Rao-Scott chi-square tests, and the associated p-values. Our example will focus on the bivariate analysis

of selected characteristics and smoking status (current smoker vs. non-smoker) using the SAS SurveyFREQ procedure.

Preparatory work to define datasets with relevant statistics

To determine the output components needed from the SurveyFreq, we use the ODS trace statements to see what SAS dataset names are provided for the objects generated by SAS. These objects contain all the statistics printed in the output window. The ODS trace statement is used twice. The "ODS trace on" statement turns the output tracing on and the "ODS trace off " statement turns off the output tracing. The SAS SurveyFreq code with the ODS trace is provided as follows.

```
ODS trace on;
Proc SurveyFREQ  data=mmp nomcar;
table basepop * _agegrp3*_cursmoker / row cl cv chisq1;
strata nat_strat_owt;
cluster nat_clust_owt;
weight nat_owt;
format _agegrp3 agegr4p.;
format basepop _yesno.;
run;
ODS trace off;
```

When this code is run, the partial listing resulting from ODS trace statements in the log window is shown below. Two datasets are created, CrossTabs and ChiSq1. The log also tells us what statistics are included: the Crosstabs dataset contains all of the statistics generated from the crosstabs in the table statement, including the unweighted and weighted frequencies, row and column percentages, standard error, 95% confidence interval, and coefficient of variation; the ChiSq1 dataset contains the results of the chi-square test, specifically, the value of the modified Rao-Scott chi-square and the p-value that is associated with the chi-square statistic.

```
Output Added:
-------------
Name:        CrossTabs
Label:       CrossTabulation Table
Template:    Stat.SurveyFreq.CrossTabFreqs
Path:        Surveyfreq.Table1of1.CrossTabs
-------------

Output Added:
-------------
Name:        ChiSq1
Label:       Rao-Scott Modified Chi-Square Test
Template:    Stat.SurveyFreq.Factoid
Path:        Surveyfreq.Table1of1.ChiSq1
-------------
```

A partial listing of the crosstabs dataset below shows the statistics that are included in the dataset.

F__AGEGRP3	F__cursmoker	Frequency	WgtFreq	RowPercent	RowStdErr	RowLower CL	RowUpper CL	RowCV
18-29	No	192	19289	62.4471	2.5916	57.3381	67.556	0.0415
18-29	Yes	123	11599	37.5529	2.5916	32.444	42.6619	0.069
18-29	Total	315	30888	100	_	_	_	_
30-39	No	406	41537	57.5709	2.71	52.2287	62.9131	0.0471
30-39	Yes	316	30613	42.4291	2.71	37.0869	47.7713	0.0639
30-39	Total	722	72150	100	_	_	_	_

……

Total	No	2427	241965	_	_	_	_	_
Total	Yes	1780	177980	_	_	_	_	_
Total	Total	4207	419945	_	_	_	_	_

Step 1: Create the SAS macro (I): SURVEYFREQ procedure

We generalize the above SURVEYFREQ procedure into a macro so that the statistics table can be generated for any variable without the need of repeated typing all of the above SAS code. The macro uses the following arguments: the base population variable (basepop), the row characteristic variable being analyzed (rowvar), the column or outcome variable (colvar), the name of the output crosstab dataset (cross), the name of the output p-value dataset (pvalue), and the characteristic string describing the variable (char, for table reporting purpose). Here is how the macro is coded:

```
%macro responsetable (basepop,colvar,rowvar,cross, pvalue, char);

proc surveyfreq data=mmp nomcar;
table &basepop*&rowvar*&colvar / row cl cv chisq1;
strata nat_strat_owt;
cluster nat_clust_owt;
weight nat_owt;
ods output crosstabs=&cross chisq1=&pvalue;
run;
```

Step 2: Create the SAS macro (II): Organize the SURVEYFREQ output

In this step, we create two new datasets which are produced from reading in the outputs from PROC SURVEYFREQ. In the first dataset, the keep statement is used to retain the desired statistics for the specific analyses and the where statement is used to specify any restrictions to the data that will be included in the table. In this example, the keep statement is used to retain all of the desired statistics - the unweighted n (frequency), row % (rowpercent), 95% ci (upper and lower confidence limits) and coefficient of variation (cv) for each of the variables in the analyses (age, race, etc.). The where statement specifies that the analysis is restricted to basepop =1 and identifies that totals should be removed from the tables to be produced. In this analysis, “Total” and non-base population rows are excluded from the dataset because they are not used for reporting. In the second dataset, we keep the p-value of the chi-square statistic.

```
data &cross ;
length F_&rowvar $90 characteristic $195;
format percent lowerCL upperCL CV 8.1;
set &cross;
where &basepop=1  and F_&rowvar ne "Total";
characteristic=&char;
keep characteristic F_&colvar F_&rowvar &rowvar frequency percent lowerCL upperCL CV;
rename F_&rowvar= level;
run;

proc sort data=&cross ;
by &var;
run;

data &pvalue;
set  &pvalue;
where Label1 contains "Pr > ChiSq";
characteristic=&char;
keep  characteristic cvalue1;
rename cvalue1=pvalue;
run;

%mend;
```

Step 3. Generate the statistics datasets from the SAS procedures

In this step, the SAS macro is invoked to generate an output and save all needed parameters for each characteristic variable (age, race, etc.) needed for the table shell. The arguments used in the macro are filled for different characteristic variables each time the macro is invoked. An example of the SAS code to invoke the macro for age group (_agegrp3) is provided below. This statement will generate a table of basepop * _agegrp3*_smoker. This part of the SAS code will also save the outputs into two datasets, crosstable1 and ptable1. Each of these datasets will have the statistics described above for each level of age group by smoker categories.

```
%responsetable (BasePop, _cursmoker, _agegrp3, crosstable1, ptable1, 'Age');
```

Step 4. Organize the results obtained from invoking the macro

This step reads in the datasets created for each characteristic variable using the macro as described in Step 3, combines the datasets into one, and manipulates how the data are presented. In the dataset, the 95% confidence interval is originally presented in two numeric variables: lower limit value and upper limit value. In this step, SAS functions commands are used to produce the 95% confidence interval in the format of (lower value – upper value). Also, the results from the cross-tabulations are merged with that from the p-value. This step creates one dataset with all values, and the structure of the dataset makes it ready for populating the final reporting table. An example of the macro and a statement to invoke the macro that creates the new formatted confidence level is as follows.

```
%macro treat (dataset,varseq);
data &dataset;
set  &dataset;
ci&varseq=cat(put(percent,8.1),' (',strip(put(lowerCL,8.1)),' - ',strip(put(upperCL,8.1)),')');
keep characteristic level frequency ci&varseq;
rename frequency=f&varseq;
run;
%mend;

%treat (col1,1);
```

Step 5. Use Proc Report to output the tables

The Proc report will use the data created in the previous step and format the output. The multiple STYLE options in the PROC REPORT statement define the layout and structure of the table, including the font used in the title and column and row headers, the frame and separating lines for the table, the background color, the dimension of each column and cell, the alignment of the texts, and so on.

```
proc report data=table split='*' nowindows style(column)={asis=on}
style(report)={font face="times new roman" font size=12pt rules=groups frame=hsides
               background=white borderwidth = .01cm posttext=" "}
style(column)={font_face="times new roman" just=left cellspacing = 0 font_size=1.5
               cellwidth=.6in rules=none frame=void background=white}
style(header)={font face="times new roman" font weight=bold just=center vjust=bottom
               cellheight=0.3in font_size=10pt cellwidth=.6in rules=rows
               frame=hsides background=white bordercolor=black}
style(lines) ={};

column characteristic level ('Total' f1 ci1) ('Smoker' f2 ci2) ('Non-smoker' f3 ci3)  pvalue;

define characteristic / group noprint order=data width=60 flow;

define level / display ' ' CENTER order=internal
                       style(column)={cellwidth=2in just=left}
                       style(header)={cellwidth=2in just=left};
```

The define statements describe the presented variables and their usages. For the variable called “characteristic”, we use the NOPRINT option. Used along with the following COMPUTE statement, it can create subtitles (such as “age”, “gender”, etc.) in the first column shown in table shell.

```
compute before characteristic / style = [font_weight=bold];
line @1 characteristic $195. ;
endcomp;
```

We also use the ODS output statement to print the reporting table into an .rtf file, and the table can therefore be permanently saved after the SAS session is ended.

Summary

Weighted analysis of national survey data tends to result in a large quantity of summary statistics using the PROC SURVEYFREQ procedure. Utilizing the ODS technique enables the selection, sorting, and storage of the important pieces of information from the huge resource of SAS output.

The presented PROC REPORT code can be readily used to generate a large amount of summary tables for publication use. It is, on the other hand, also flexible enough to accommodate individualized table structure and format requirements. The full code for the production of publication-ready tables for survey data can be found in Appendix A.

References

Blair J, McNaghten A, Frazier E, Skarbinski J, Huang P, Heffelfinger J. Clinical and behavioral characteristics of adults receiving medical care for HIV infection—Medical Monitoring Project, United States, 2007. MMWR Surveill Summ. 2011;60(SS-11):1-20.

Frankel M, McNaghten A, Shapiro M, et al. A probability sample for monitoring the HIV-infected population in care in the U.S. and in selected states. Open AIDS J. 2012; 6(suppl 1):67-76. doi:10.2174/1874613601206010067.

ACKNOWLEDGEMENTS

We would like to thank the Clinical Outcomes Team for the challenges and opportunities to work to produce these tables over the past years.

CONTACT INFORMATION

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

Names: Emma Frazier or Shuyan Zhang
Enterprise: Centers for Disease Control and Prevention
Address: 1600 Clifton Rd, MS K-30
City, State, ZIP: Atlanta, GA 30341
Email: elf3@cdc.gov (Emma Frazier) or vzk3@cdc.gov (Shuyan Zhang)

Appendix A - Full Code with Documentation

```
*****************************************************************************************;
*                 SAS macro to produce publication ready tables                          ;
* -------------------------------------------------------------------------------------- ;
*  The purpose of this macro is to simplify the process of analyzing survey data for     ;
*  different variables and to output the statistics into datasets for creating report    ;
*  tables. The macro is useful for researchers who need to produce a large amount of     ;
*  publication-ready tables resulting from SAS Survey procedures.                        ;
*****************************************************************************************;

%macro responsetable (basepop,colvar,rowvar,cross, pvalue, char);

*****************************************************************************************;
* Step 1:  SURVEYFREQ procedure                                                          ;
* -------------------------------------------------------------------------------------- ;
*  In this step, we use the SURVEYFREQ procedure to make the weighted analysis on the data ;
*  and save the resulting crosstabs and other statistics into datasets.                  ;
*****************************************************************************************;
proc surveyfreq data=mmp nomcar;
table &basepop*&rowvar*&colvar / row cl cv chisq1;
strata nat_strat_owt;
cluster nat_clust_owt;
weight nat_owt;
ods output crosstabs=&cross chisq1=&pvalue;
run;

*****************************************************************************************;
* Step 2:  Organize the  SURVEYFREQ output                                               ;
* -------------------------------------------------------------------------------------- ;
*  In Step 2, we select the needed statistics from the output datasets and organize them  ;
*  for table creation                                                                    ;
*****************************************************************************************;
data &cross ;
length F_&rowvar $90 characteristic $195;
format percent lowerCL upperCL CV 8.1;
set &cross;
where &basepop=1  and F_&rowvar ne "Total";
characteristic=&char;
keep characteristic F_&colvar F_&rowvar &rowvar frequency percent lowerCL upperCL CV;
rename F_&rowvar= level;
run;

proc sort data=&cross ;
by &rowvar;
run;

data &pvalue;
set  &pvalue;
where Label1 contains "Pr > ChiSq";
characteristic=&char;
keep  characteristic cvalue1;
rename cvalue1=pvalue;
run;

%mend;


*****************************************************************************************;
* Step 3.  Generate the statistics datasets by invoking the macro                        ;
* -------------------------------------------------------------------------------------- ;
*  In this step, the SAS macro is invoked to generate an output and save all selected    ;
*  parameters for each characteristic variable (age, race, etc.)needed for the table     ;
*  shell.                                                                                ;
*****************************************************************************************;

%responsetable (BasePop, _cursmoker, _agegrp3, crosstable1, ptable1, 'Age');
%responsetable (BasePop, _cursmoker, _gender,  crosstable2, ptable2, 'Gender');
```

```
%responsetable (BasePop, _cursmoker, _newrace2,  crosstable3, ptable3, 'Race/Ethnicity');
%responsetable (BasePop, _cursmoker, _educ,  crosstable4, ptable4, 'Education');

proc surveyfreq data=mmp nomcar;
table basepop*_agegrp3*_cursmoker / row cl cv chisq1;
strata nat_strat_owt;
cluster nat_clust_owt;
weight nat_owt;
ods output crosstabs=cross1 chisq1=pvalue1;
run;


*************************************************************************************;
* Step 4.  Organize the results obtained from invoking the macro                    ;
* ---------------------------------------------------------------------------------- ;
*  This step reads in the datasets created for each characteristic variable using the macro;
*  as described in Step 3, combines the datasets into one, and manipulates how the data are;
*  presented.                                                                        ;
*************************************************************************************;

data col1 col2 col3;
retain F__cursmoker characteristic level  frequency percent lowerCL upperCL CV;
set crosstable1 - crosstable4;
keep F__cursmoker characteristic level  frequency percent lowerCL upperCL CV;
if F__cursmoker='Total' then output col1;
if F__cursmoker='Yes'  then output col2;
if F__cursmoker='No'   then output col3;
run;

%macro treat (dataset,varseq);
data &dataset;
set  &dataset;
ci&varseq=cat(put(percent,8.1),' (',strip(put(lowerCL,8.1)),' - ',strip(put(upperCL,8.1)),')');
keep characteristic level frequency ci&varseq;
rename frequency=f&varseq;
run;
%mend;

%treat (col1,1);
%treat (col2,2);
%treat (col3,3);


data freqtable;
set col1;
set col2;
set col3;
level=cat(' ',level);
order=_N_;
run;

data pvaltable;
length pvalue $8 characteristic $195;
set ptable1-ptable4;
run;

proc sort data=freqtable; by characteristic; run;
proc sort data=pvaltable; by characteristic; run;

data table;
merge freqtable pvaltable; by characteristic;
if not first.characteristic then do; pvalue=''; end;
run;

proc sort data=table out=table (drop=order); by order; run;

*************************************************************************************;
* Step 5.  Use Proc Report to output the tables                                     ;
* ---------------------------------------------------------------------------------- ;
*  In this step, Proc Report uses the data created in the previous steps and format  ;
*  the output according to the desired table structure.                              ;
```

```
*************************************************************************************************;

title;
options nodate orientation=landscape nonumber;
ods escapechar='^';
ods rtf file = "&out\table.rtf" ;
proc report data=table split='*' nowindows style(column)={asis=on}
style(report)={font_face="times new roman" font_size=12pt rules=groups frame=hsides
               background=white borderwidth = .01cm posttext=" "}
style(column)={font_face="times new roman" just=left cellspacing = 0 font_size=1.5
               cellwidth=.6in rules=none frame=void background=white}
style(header)={font face="times new roman" font weight=bold just=center vjust=bottom
               cellheight=0.3in font_size=10pt cellwidth=.6in rules=rows
               frame=hsides background=white bordercolor=black}
style(lines) ={};

ODS TEXT='^S={LEFTMARGIN=0.9in RIGHTMARGIN=0.9in fontsize=12pt fontweight=bold font_face="times
new roman"}
Table1. Prevalence of smoking among HIV-infected persons in care, by smoking status - Medical
Monitoring Project, United States, 2009';
ODS TEXT='^S={LEFTMARGIN=0.9in RIGHTMARGIN=0.9in fontsize=12pt font_face="times new roman"}';

column characteristic level ('Total' f1 ci1) ('Smoker' f2 ci2) ('Non-smoker' f3 ci3)  pvalue;

define characteristic / group noprint order=data width=60 flow;

define level / display ' ' CENTER order=internal
                        style(column)={cellwidth=2in just=left}
                        style(header)={cellwidth=2in just=left};
define f1    /  'N' CENTER flow
                        style(column)={cellwidth=0.6in just=right}
                        style(header)={cellwidth=0.6in just=center};
define ci1   /  '%(95% CI)' CENTER WIDTH=25
                        style(column)={cellwidth=1.5in just=right}
                        style(header)={cellwidth=1.5in just=center};
define f2    /  'N' CENTER flow
                        style(column)={cellwidth=0.6in just=right}
                        style(header)={cellwidth=0.6in just=center};
define ci2   /  '%(95% CI)' CENTER WIDTH=25
                        style(column)={cellwidth=1.5in just=right}
                        style(header)={cellwidth=1.5in just=center};
define f3    /  'N' CENTER flow
                        style(column)={cellwidth=0.6in just=right}
                        style(header)={cellwidth=0.6in just=center};
define ci3   /  '%(95% CI)' CENTER WIDTH=25
                        style(column)={cellwidth=1.5in just=right}
                        style(header)={cellwidth=1.5in just=center};
define pvalue/  'p-vlaue' CENTER WIDTH=25
                        style(column)={cellwidth=1.5in just=center}
                        style(header)={cellwidth=1.5in just=center};

compute before characteristic / style = [font_weight=bold];
line @1 characteristic $195. ;
endcomp;
run;

ods rtf close;
```