Paper PH-86

You've used FREQ, but have you used SURVEYFREQ?

Charlotte Baker, Florida Agricultural and Mechanical University

ABSTRACT

PROC FREQ is a well utilized procedure for descriptive statistics. If the data being analyzed is from a complex survey sample, it is best to use the PROC SURVEYFREQ procedure instead. Other than the SAS® documentation on PROC SURVEYFREQ, few user examples exist for how to perform analyses using this procedure. This paper will demonstrate why PROC SURVEYFREQ should be used and how to implement it in your research.

INTRODUCTION

Calculating descriptive statistics is one of the first steps when familiarizing yourself with a new data set, cleaning data, or conducting analyses. The frequency is an important summary measure that is typically used for assessing categorical data and continuous data that does not have many unique values. We may want to find out how many observations meet certain criteria (the frequency), identify associations between two or more variables with a chi-square statistic, or find out more about a subset of the population. PROC FREQ is the first tool that many SAS® users learn to do these things. However, if the data was created with multi-stage probability sampling and is intended to be used as such, PROC SURVEYFREQ is the better tool.

Multi-stage probability sampling is the combined use of various sampling methods, including cluster sampling and stratified sampling, to collect information on multiple segments of a population in order to ensure truly random but representative selection. Weights can be applied to this data to provide regional or national estimates while taking into account known distributions of age, sex, race, etc. For this paper, we will refer to this entire sampling process as complex survey sampling. Many national health surveys, such as the Behavioral Risk Factor Surveillance System (BRFSS) and the Healthcare Cost Utilization Project Kids Inpatient Database (HCUP KID), utilize complex survey sampling to produce data that represents the health status of the entire United States.

This paper will discuss some considerations for writing and running PROC SURVEYFREQ code, and important output to pay attention to and interpret as you implement PROC SURVEYFREQ in your research.

FREQUENCIES FOR A COMPLEX SURVEY

To obtain the best results for frequencies using PROC SURVEYFREQ, we begin with the complete original data set. This data set should include variables representing strata, weights, and clusters that were utilized in the sampling. Some data sets may not include all three of these, but if they are present they should be taken into account. Using the entire set ensures that any descriptive statistics or other analysis will appropriately take advantage of the sampling design.

DIFFERENCES IN THE PROC SURVEYFREQ AND PROC FREQ CODE

The basic structure of PROC SURVEYFREQ code has some similarities to the PROC FREQ, but also has several key differences.

```
PROC FREQ <options> ;
BY variables ;
EXACT statistic-options </ computation-options> ;
OUTPUT <OUT=SAS-data-set> options ;
TABLES requests </ options> ;
TEST options ;
WEIGHT variable </ option> ;
RUN;

PROC SURVEYFREQ <options>;
BY variables ;
CLUSTER variables ;
REPWEIGHTS variables </ options> ;
STRATA variables </ option> ;
TABLES requests </ options> ;
WEIGHT variable ;
RUN;
```

The differences are highlighted in yellow above. As we have already discussed, PROC SURVEYFREQ takes into account sampling clusters and strata that PROC FREQ cannot. This is the primary reason for using PROC SURVEYFREQ instead of using PROC FREQ. The only required statements for either procedure are the PROC statements and RUN. It is good practice to specify what data set the procedures are to use. If more than one-way tables are necessary or we only need information on specific variables, the TABLES statement is also required. For PROC SURVEYFREQ, there can only be one WEIGHT statement. However, multiple TABLES, STRATA, REPWEIGHTS, or CLUSTER statements can be used in the same PROC step. The STRATA, CLUSTER, and WEIGHT statements are only necessary if the data contains strata, clusters, or weights respectively. If using replicate weighting, no STRATA or CLUSTER statements are needed.

EXAMPLES

The three examples in this paper are based on the following scenario. We are conducting a study using the 2012 BRFSS data from the Centers for Disease Control and Prevention (CDC). All data used in these examples is freely available and the location of the data can be found in the References section of this paper.

The following three examples demonstrate three situations in which to use PROC SURVEYFREQ for survey data. All examples use the 2012 BRFSS data from the Centers for Disease Control and Prevention. This data is freely available and the location of the data can be found in the References section of this paper. All three examples work through the basic steps of a conducting a study that is meant to describe the relationship between diabetes and arthritis among sedentary people in the United States in 2012. For this study, the alpha level for significance testing is 0.05.

According to the CDC, the variable _LLCPWT should be used as the weight variable, _STSTR should be used as the stratification variable, and _PSU should be used as the cluster variable. We will use the variable _DRDXAR1 for arthritis, modify the variable DIABETE3 for diabetes, and modify the variable _TOTINDA for sedentary lifestyle. The modifications will be used only to create two level variables for this analysis.

Example 1

We are interested in finding out how many residents of the United States have ever been diagnosed with arthritis and how many have ever been diagnosed with non-gestational diabetes. We are also interested in finding out how many people have not exercised in the last 30 days (the study definition of sedentary).

If we were to use PROC FREQ and the weight variable _LLCPWT our code might look like this:

```
DATA brfss2;
SET brfss;
DIABETES = DIABETE3;
IF DIABETES in (2, 3, 4, 7) then DIABETES = 0;
IF DIABETES = 9 THEN DIABETES = .;
IF _TOTINDA = 9 THEN _TOTINDA = .;
RUN;

PROC FREQ data = brfss2 ;
TABLES _DRDXAR1 DIABETES _TOTINDA;
WEIGHT _LLCPWT;
RUN;
```

If we use PROC SURVEYFREQ, our code might look like this:

```
PROC SURVEYFREQ data = brfss2;
CLUSTER _PSU ;
STRATA _STSTR ;
TABLES _DRDXAR1 DIABETES _TOTINDA;
WEIGHT _LLCPWT ;
RUN;
```

We obtain the following results from PROC FREQ:

RESPONDENTS DIAGNOSED WITH ARTHRITIS				
_DRDXAR1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Diagnosed with arthritis	61966255	25.62	61966255	25.62
Not diagnosed with arthritis	1.7993E8	74.38	2.419E8	100.00

Frequency Missing = 1158411.6861

(EVER TOLD) YOU HAVE NON-GESTATIONAL DIABETES				
DIABETES	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No/Don't Know	2.1819E8	89.82	2.1819E8	89.82
Yes	24727034	10.18	2.4292E8	100.00

Frequency Missing = 139715.80742

LEISURE TIME PHYSICAL ACTIVITY CALCULATED VARIABLE				
_TOTINDA	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Had physical activity or exercise	1.8429E8	76.48	1.8429E8	76.48
No physical activity or exercise in last 30 days	56681461	23.52	2.4097E8	100.00

Frequency Missing = 2084308.344

Output 1. Arthritis, Diabetes, and Sedentary Lifestyle Weighted Frequency Output from PROC FREQ

We obtain the following results from PROC SURVEYFREQ:

Data Summary	
Number of Strata	1102
Number of Clusters	475687
Number of Observations	475687
Sum of Weights	243057710

RESPONDENTS DIAGNOSED WITH ARTHRITIS					
_DRDXAR1	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Diagnosed with arthritis	162758	61966255	292022	25.6166	0.1204
Not diagnosed with arthritis	310423	179933043	494190	74.3834	0.1204
Total	473181	241899298	464580	100.000	
Frequency Missing = 2506					

(EVER TOLD) YOU HAVE NON-GESTATIONAL DIABETES					
DIABETES	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
No/Don't Know	415665	218190960	478665	89.8208	0.0868
Yes	59763	24727034	213525	10.1792	0.0868
Total	475428	242917994	465751	100.000	
Frequency Missing = 259					

LEISURE TIME PHYSICAL ACTIVITY CALCULATED VARIABLE					
_TOTINDA	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Had physical activity or exercise	355588	184291941	470669	76.4781	0.1304
No physical activity or exercise in last 30 days	118540	56681461	333420	23.5219	0.1304
Total	474128	240973402	462176	100.000	
Frequency Missing = 1559					

Output 2. Arthritis, Diabetes, and Sedentary Lifestyle Weighted Frequency Output from PROC SURVEYFREQ

While the weighted frequencies obtained from PROC FREQ (Output 1) and PROC SURVEYFREQ (Output 2) appear to be equal, we know those obtained from PROC SURVEYFREQ are more accurate for the nationally weighted estimate because they take into account the strata and clusters. The “Sum of Weights” in the first table of results from PROC SURVEYFREQ gives the total weighted population size.

Example 2

We now want to look at the relationship between diabetes and arthritis for the entire population before we focus on just those that are sedentary.

Our PROC SURVEYFREQ code might look like this:

```
PROC SURVEYFREQ data = brfss2;
CLUSTER _PSU ;
STRATA _STSTR ;
TABLES _DRDXAR1*DIABETES ;
WEIGHT _LLCPWT ;
RUN;
```

From the above code, we obtain the following results:

Table of _DRDXAR1 by DIABETES						
_DRDXAR1	DIABETES	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Diagnosed with arthritis	No/Don't Know	129756	49679462	266392	20.5451	0.1102
	Yes	32954	12269198	143423	5.0740	0.0592
	Total	162710	61948660	291987	25.6191	0.1205
Not diagnosed with arthritis	No/Don't Know	283832	167509395	491607	69.2741	0.1309
	Yes	26483	12348532	163045	5.1068	0.0668
	Total	310315	179857927	494134	74.3809	0.1205
Total	No/Don't Know	413588	217188857	477373	89.8192	0.0871
	Yes	59437	24617730	213303	10.1808	0.0871
	Total	473025	241806587	464527	100.000	
Frequency Missing = 2662						

Output 3. Output from Crosstabulation of Arthritis and Diabetes in PROC SURVEYFREQ

Unlike the default PROC FREQ output, PROC SURVEYFREQ does not provide a 2x2 table with row and column percentages for the results. However, the default output for PROC SURVEYFREQ does include the weighted frequencies and percentages for each group so one can obtain the same data as needed.

The results show some overlap between arthritis and diabetes in our BRFSS data. There are people who have arthritis and diabetes. We want to see if this relationship is statistically significant. For this, we use the chi-square test statistic. We add the CHISQ option to the TABLES statement just as we would with PROC FREQ. The CHISQ option in PROC FREQ provides a Pearson chi-square value by default while the CHISQ option in PROC SURVEYFREQ provides a Rao-Scott chi-square value by default.

```
PROC SURVEYFREQ data = brfss2;
CLUSTER _PSU ;
STRATA _STSTR ;
TABLES _DRDXAR1*DIABETES / CHISQ ;
WEIGHT _LLCPWT ;
RUN;
```

The following table (the third table of results) is found under the crosstabulation table when the CHISQ option is requested:

Rao-Scott Chi-Square Test	
Pearson Chi-Square	16504.7681
Design Correction	3.8375
Rao-Scott Chi-Square	4300.8712
DF	1
Pr > ChiSq	<.0001
F Value	4300.8712
Num DF	1
Den DF	471923
Pr > F	<.0001
Sample Size = 473025	

Output 4. Chi-Square Results from Crosstabulation of Arthritis by Diabetes in PROC SURVEYFREQ

We see that the Rao-Scott chi-square value 4300.8712 has a p-value of <.0001. Because that p-value is less than our designated alpha level of 0.05, we say that there is a significant relationship between diabetes and arthritis in the United States population.

Example 3

Now that we have evaluated our variables individually and looked at the cross-tabulation results of arthritis by diabetes, we are ready to answer our main research question. We want to know if there is a relationship between diabetes and arthritis among sedentary people in the United States. We can find an unadjusted answer using PROC SURVEYFREQ.

If we did not have data from complex survey sampling, we might create a new data set that only contains sedentary people. Because we do have this type of data, we stratify the data instead so that we can continue taking advantage of the complex survey sampling structure. Our variable for being sedentary (_TOTINDA) already has two levels. If it did not, we would create a variable for sedentary that did have two levels and use it for this analysis. Instead of using a BY statement, we create a three-way table in order to take advantage of the complex survey sample. We continue to ask for the chi-square test statistic so that we can see if any relationship is significant.

```
PROC SURVEYFREQ data = brfss2;
CLUSTER _PSU ;
STRATA _STSTR ;
TABLES _TOTINDA*_DRDXAR1*DIABETES / CHISQ ;
WEIGHT _LLCPWT;
RUN;
```

In the output, we look for the specific results of interest. In this case we are looking for the results that begin with "_TOTINDA = 0" (or "_TOTINDA=No physical activity or exercise in last 30 days" if you are using the BRFSS provided formats).

Table of _DRDXAR1 by DIABETES						
Controlling for _TOTINDA=No physical activity or exercise in last 30 days						
_DRDXAR1	DIABETES	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Diagnosed with arthritis	No/Don't Know	39995	15153397	157695	26.9109	0.2588
	Yes	14926	5420019	94283	9.6254	0.1635
	Total	54921	20573416	181417	36.5363	0.2886
Not diagnosed with arthritis	No/Don't Know	54141	31916161	280372	56.6799	0.3061
	Yes	8651	3819904	89920	6.7838	0.1554
	Total	62792	35736064	291870	63.4637	0.2886
Total	No/Don't Know	94136	47069558	313711	83.5908	0.2175
	Yes	23577	9239923	129418	16.4092	0.2175
	Total	117713	56309481	332281	100.000	

Rao-Scott Chi-Square Test	
Pearson Chi-Square	4877.1516
Design Correction	4.0441
Rao-Scott Chi-Square	1205.9972
DF	1
Pr > ChiSq	<.0001
F Value	1205.9972
Num DF	1
Den DF	470390
Pr > F	<.0001
Sample Size = 471492	

Output 5. Selected Three Way Table Output from PROC SURVEYFREQ

The results of this PROC SURVEYFREQ indicate that for sedentary people there is a significant relationship between diabetes and arthritis (Rao-Scott chi-square 1205.9972, p <.0001). We do not have to refer to the results that begin with "_TOTINDA=1" as they are not of interest for this analysis.

COMPUTING CONSIDERATIONS

When considering using PROC SURVEYFREQ for health data we have to consider the available computing power including the amount of RAM available. In our experience, this is less of an issue when attempting to use PROC SURVEYFREQ on small data sets (less than 10000 observations). However, if using larger datasets, running analyses can sometimes slow the computer down and may indeed prevent the usage of the computer while analyses

are running, particularly for situations such as example three above. Greater processing speed and RAM on a single computer or availability of a remote server to run analyses may be necessary.

CONCLUSION

PROC SURVEYFREQ is a tool that should be in every programmers toolbox if they utilize data that comes from complex survey sampling techniques. It allows the incorporation of the strata, replicate weights, and clusters that cannot be taken into account with PROC FREQ. PROC SURVEYFREQ is capable of assisting anyone with calculating descriptive statistics for categorical data and some continuous data prior to completing other analyses. Knowing how to use PROC SURVEYFREQ and other PROC SURVEY techniques eliminates the need for using alternative software for survey analysis.

REFERENCES

- SAS Institute Inc. SAS® 9.3 Help and Documentation. Cary, NC: SAS Institute Inc., 2014.
- Centers for Disease Control and Prevention. BRFSS 2012 Survey Data and Documentation, 2013, Available at http://www.cdc.gov/brfss/annual_data/annual_2012.html

CONTACT INFORMATION

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

Charlotte Baker
Florida A&M University Institute of Public Health
1515 S. Martin Luther King, Jr Blvd
Tallahassee, FL 32307
charlotte.m.h.baker@gmail.com