Paper 116

Merging and Analysis of Complex Survey Data Sets by using Proc Survey Procedures in SAS®

Nushrat Alam, Masters of Public Health (Student), Institute of Public Health, Florida Agricultural and Mechanical University

ABSTRACT:

This paper is focused on merging and analysis of the complex survey data sets. The sample design of any complex survey data is consists of stratification, clustering, multi-stage sampling, and unequal probability of selection of observations. This paper provides an outline of merging of different complex survey datasets and the use multiple SAS® procedures like PROC SURVEYMEANS, PROC SURVEYFREQ, and PROC SURVEYREG to analyze different variables.

INTRODUCTION:

Survey is one of the most popular methods of collecting data from general population. In public health study, survey remains the most commonly used tool to collect any health related data from any identified sample population. Selecting samples from population and analysis of this survey data follow some defined procedure. Survey data requires special attention while performing descriptive and quantitative statistical analysis. The simple random sampling (SRS) method gives all the samples of any population independently equal probability of being selected for analysis. (Lee & Forthofer, 2006). Conducting surveys following the SRS is often expensive, time consuming and practically impossible to perform. In real world practice, collecting samples for any national census is more complicated and need multilevel approach. For this, a complex survey design is often applied to collect the most representative sample of the population. Consumer Expenditure Survey, Behavioral Risk Factor Surveillance System (BRFSS), National Survey of Family Growth (NSFG) are some of the examples of complex survey.

NHANES DATA SETS:

The data sets used for this paper were selected from National Health and Nutrition Examination Survey (NHANES) 2009-2010. The samples were collected from the specific United States population by following a complex, multistage probability sampling design. Stratification and clustering are the steps responsible for the complexity of any given complex survey design ("Analysing Complex Survey Data: Clustering, Stratification and Weights," n.d.). The first stage of sampling these datasets was to define the strata which are the small groups of counties. These groups are also known as primary sampling unit (PSU) for this survey. The next steps were to selecting specific segments of the PSUs containing the clusters of the households. The last two steps follow the selection of the households and selection of the individuals from those households. ("Vital and Health Statistics Report Series 1, Number 56 August 2013 - sr01_056.pdf," n.d.). The raw sample of any national survey contains samples that are not always representing the population. In order to use those sample appropriately a process called weighting in done. It is applied to remove the unequal probability and nonresponse from the samples (Lee & Forthofer, 2006). The main purpose of weighting is make the sample free from any bias and balancing the under and over representativeness if any subgroup of population ("Analysing Complex Survey Data: Clustering, Stratification and Weights," n.d.). The weight used for the survey was mentioned as "wtint2yr".

SAS Procedures Used in Analyzing Complex Survey Design:

A large portion of these survey data analysis are performed by using different versions of SAS. SAS has some unique functions dedicated for analyzing any survey-design featuring stratification, multilevel clustering, and probability sampling weight (Siller & Tompkins, n.d.). Listed below are some of the examples of different SAS procedures:

- PROC SURVEYMEANS
- PROC SURVEYFREQ
- PROC SURVEYREG

This paper will focus on the procedures mentioned earlier along with merging different datasets of same survey using SAS DATA step.

MERGING DATASETS:

The DATA step in SAS is used to merge two or more datasets together. Datasets can be merged by using common variables, values or groups. This is called Match Merging. Before merging any datasets the first step is to sort those datasets by using the common variables. PROC SORT and BY statement are used to perform this procedure. The example datasets used in this paper had a common variable named SEQN. PROC SORT DATA = option names the datasets to be sorted. The BY= options names the common variable used for merging the datasets. Four different datasets contained information about Body Mass Index (bmx_f), diabetes (diq_f), oral health condition (ohxref_f) and demographic (demo_f) were merged together in one dataset named merge.

The source code:

```
proc sort data=sesug.bmx_f; by SEQN; run;
proc sort data=sesug.diq_f; by SEQN; run;
proc sort data=sesug.ohxref_f; by SEQN; run;
proc sort data=sesug.demo_f; by SEQN; run;


 data sesug.merge;
 merge sesug.bmx_f sesug.diq_f sesug.ohxref_f sesug.demo_f;
      run;
```

Sesug is the library name. The DATA and MERGE step do not give any visual presentation of the new dataset merge. So, to check for the result we can always check the Log statement or add another statement as PROC CONTENTS.

The Source code:

```
proc contents data=sesug.merge varnum;
      run;
```

The Output:

PROC CONTENTS statement presented the summary of the new dataset merge as well as the list of the variables the merge dataset contained as below:

The SAS System

The CONTENTS Procedure

Data Set Name	SESUG.MERGE	Observations	10537
Member Type	DATA	Variables	95
Engine	V9	Indexes	0
Created	07/25/2015 15:48:43	Observation Length	760
Last Modified	07/25/2015 15:48:43	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Figure 1. Result of the PROC CONTENTS Step

Variables in Creation Order						
#	Variable	Type	Len	Format	Informat	Label
1	SEQN	Num	8			Respondent sequence number
2	BMDSTATS	Num	8			Body Measures Component Status Code
3	BMXWT	Num	8			Weight (kg)
4	BMIWT	Num	8			Weight Comment
5	BMXRECUM	Num	8			Recumbent Length (cm)
6	BMIRECUM	Num	8			Recumbent Length Comment
7	BMXHEAD	Num	8			Head Circumference (cm)
8	BMIHEAD	Num	8			Head Circumference Comment
9	BMXHT	Num	8			Standing Height (cm)
10	BMIHT	Num	8			Standing Height Comment

Figure 2. The table with the list of the variables in the merge dataset (Partial Table)

Merging different datasets can also be done without the BY statement. In those cases, SAS combines the first observation in all data sets mentioned in the MERGE statement into the first observation in the new data set, the second observation in all data sets into the second observation in the new data set, and so on. This procedure is known as One-to-One Merging ("Merging SAS Data Sets: Match-Merging :: Step-by-Step Programming with Base SAS(R) Software," n.d.).

PROC SURVEYMEANS:

PROC SURVEYMEANS procedure is used to produce descriptive statistics for survey sample. For a complex survey design it is important to include the sampling method in the PROC SURVEYMEANS steps to accomplish the accurate results. For example, the survey data for this paper was collected following a complex stratification and clustering method. Without mentioning the strata and cluster in STRATA and CLUSTER steps the PROC SURVEYMEANS will follow simple random sampling method. For example:

The source code:

```
Title1 "Diabetes Recommendation and Visits to Diabetes Specialist";
title2 "Estimating Sample Mean without Strata and Cluster";
proc surveymeans data=sesug.clean total=10537 missing mean stderr;
      var diq230; /*diq230=how long ago saw a diabetes specialist*/
      weight wtint2yr;
      run;
title;
```

The PROC SURVEYMEANS= options names the dataset to be analyzed, TOTAL= options names the total the observations. This is used in the procedure to count for the variance estimates for the effect of sampling from the finite population ("SAS/STAT 9.2 User's Guide: The SURVEYMEANS Procedure (Book Excerpt) - statugsurveymeans.pdf," n.d.).

The Output:

Diabetes Recommendation and Visits to Diabetes Specialist
Estimating Sample Mean without Strata and Cluster

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	10537
Sum of Weights	301943719

Statistics			
Variable	Label	Mean	Std Error of Mean
DIQ230	How long ago saw a diabetes specialist	2.873440	0

Figure: 3 PROC SURVEYMEANS without strata and cluster (Partial Table)

On the other hand, specifying strata and cluster in PROC SURVEYMEANS procedure will give the results while considering the survey design. By default, PROC SUREYMEANS procedure uses Taylor Series method for sampling error. But VARMETHOD= options can also be used to specify other method like Balance Repeated Replication (BRB) and Jackknife.

The source Code:

```
Title1 "Diabetes Recommendation and Visits to Diabetes Specialist";
title2 "Estimating Sample Mean without FPC";
proc surveymeans data=sesug.clean missing mean stderr ;
      strata SDMVSTRA/list;
      cluster sdmvpsu;
      var diq230; /*diq230=how long ago saw a diabetes specialist*/
      weight wtint2yr;
      run;
title;
```

The procedure listed above did not include any finite population. So, in the results section it did not include the sampling rate. The standard error of mean was 0.079580.

The Output:

Statistics			
Variable	Label	Mean	Std Error of Mean
DIQ230	How long ago saw a diabetes specialist	2.873440	0.079580

Figure: 4 PROC SURVEYMEANS without Finite Population Count

Diabetes Recommendation and Visits to Diabetes Specialist
Estimating Sample Mean without FPC

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	15
Number of Clusters	31
Number of Observations	10537
Sum of Weights	301943719

Stratum Information						
Stratum Index	Masked Variance Pseudo-Stratum	N Obs	Variable	Label	N	Clusters
1	75	803	DIQ230	How long ago saw a diabetes specialist	60	2
2	76	785	DIQ230	How long ago saw a diabetes specialist	52	2
3	77	823	DIQ230	How long ago saw a diabetes specialist	57	2
4	78	829	DIQ230	How long ago saw a diabetes specialist	47	2

Figure: 5 PROC SURVEYMEANS with strata and cluster (Partial Table)

Including finite population count in the PROC SURVEYMEANS steps produce the sampling rate. The standard error of mean is also less (0.079573) for this survey.

Stratum Information								
Stratum Index	Masked Variance Pseudo-Stratum	Population Total	Sampling Rate	N Obs	Variable	Label	N	Clusters
1	75	10537	0.02%	803	DIQ230	How long ago saw a diabetes specialist	60	2
2	76	10537	0.02%	785	DIQ230	How long ago saw a diabetes specialist	52	2
3	77	10537	0.02%	823	DIQ230	How long ago saw a diabetes specialist	57	2
4	78	10537	0.02%	829	DIQ230	How long ago saw a diabetes specialist	47	2
5	79	10537	0.02%	696	DIQ230	How long ago saw a diabetes specialist	54	2

Figure: 6 PROC SURVEYMEANS with Finite Population Count (Partial Table)

PROC SURVEYFREQ:

PROC SURVEYFREQ performs the frequency procedure for any survey sample. PROC SURVEYFREQ= options names the dataset to be analyzed for frequency procedure. Additional options like NOSUMMAREY can be added in this steps. The NOSUMMERY options exclude the summary option from the frequency table. The next step is the TABLE option. Options like Confidence limit (CL), row percent (ROW), excluding total (NOTOTAL) etc. were also mentioned in the step. Frequency can be calculated by one-way or two-way tables.

For one-way table the Table option names the variables for which frequency will be counted. Statistical tests like Chi-Square test option can be added in this step to analyze the variables. For survey data analysis a modified version of Chi-Square test named Rao-Scott Chi-Square is performed. This test counts for sample design and provide results for the entire study population. Rao-Scott test first performs the Chi-Square test with the weighted value then modifies the results with a design correction. A F test value is also provided. The null hypothesis for this test reflects the equal proportion of the level of the One-Way Table (“The SURVEYFREQ Procedure: Getting Started :: SAS/STAT(R) 9.2 User’s Guide, Second Edition,” n.d.).

The Source Code:

```
Title1 "Different Races ";
Title2 "One-Way Table";
proc surveyfreq data=sesug.clean nosummary total=10537 missing;
      Tables ridreth1/cl row nototal nopercent chisq;/*ridreth1=race*/
      strata SDMVSTRA;
      cluster SDMVPSU;
      weight wtint2yr;
      run;
title;
```

The Output:

Different Races
One-Way Table

The SURVEYFREQ Procedure

Race/Ethnicity - Recode					
RIDRETH1	Frequency	Weighted Frequency	Std Dev of Wgt Freq	95% Confidence Limits for Percent	
1	2384	31665962	5895462	5.5310	15.4437
2	1133	16446175	3657823	2.7091	8.1844
3	4420	195122116	18492707	57.5010	71.7430
4	1957	36407655	3130347	10.1285	13.9871
5	643	22301811	3262564	5.1383	9.6359

Figure: 7 One-Way Frequency Table

Rao-Scott Chi-Square Test	
Pearson Chi-Square	13253.4436
Design Correction	32.2251
Rao-Scott Chi-Square	411.2765
DF	4
Pr > ChiSq	<.0001
F Value	102.8191
Num DF	4
Den DF	64
Pr > F	<.0001
Sample Size = 10537	

Figure: 8 Rao-Scott Chi-Square Test (One-Way Table)

From the table above, the F test value is 102 with a P-value of <0.001. This means we can reject the null hypothesis or in other words, the level of proportion was significantly different for the level of races.

A Two-Way table frequency was performed by using the two variables in the TABLE step. The Rao-Scott Chi-Sqaure test was requested to look for any association between gum diseases among different races.

The source code:

```
Title1 "Gum Diseasea Among Races";
Title2 "Two-Way Table";
proc surveyfreq data=sesug.clean nosummary total=10537 missing;
      Tables ridreth1*oharocgp/cl row nototal nopercent
chisq;/*ridreth1=race,OHAROCGP=Gum disease/Problem */
      strata SDMVSTRA;
      cluster SDMVPSU;
      weight wtint2yr;
      run;
title;
```

The Output:

Rao-Scott Chi-Square Test	
Pearson Chi-Square	16.3582
Design Correction	0.8683
Rao-Scott Chi-Square	18.8396
DF	8
Pr > ChiSq	0.0157
F Value	2.3550
Num DF	8
Den DF	128
Pr > F	0.0214
Sample Size = 10537	

Figure: 9 Rao-Scott Chi-Square Test (Two-Way Table)

The F test value is 2.355 with a P-value of 0.0214. This means we cannot reject the null hypothesis or in other words, the level of gum disease is not different for each level of races.

PROC SURVEREG:

The PROC SURVEYREG procedure is used to perform regression for survey data for simple survey design to any level of complex survey design. The procedure fits a linear regression model for survey data and calculate the regression coefficient together with their variance covariance matrix (“The SURVEYREG Procedure: Overview :: SAS/STAT(R) 9.22 User’s Guide,” n.d.).

The source code:

```
Title "Diabetes Recommendation, Body Mass Index and Taking Insulin";
proc surveyreg data=sesug.clean total=10537 missing;
      strata SDMVSTRA/list;
```

```
      cluster  SDMVPSU;
      class diq010;/*diq010=Doctor told you have diabetes*/
      weight wtint2yr;
      model did060 = bmxbmi  diq010 /solution;/*bmxbmi= Body Mass Index
(kg/m**2),did060=How long taking insulin*/
      run;
title;
```

The MODEL statement contains the numeric variables (dependent variable) and the CLASS statement contains the categorical variable.

The Output:

Diabetes Recommendation, Body Mass Index and Taking Insulin

The SURVEYREG Procedure

Regression Analysis for Dependent Variable DID060

Data Summary	
Number of Observations	200
Sum of Weights	5899352.1
Weighted Mean of DID060	37.40583
Weighted Sum of DID060	220670166

Design Summary	
Number of Strata	15
Number of Clusters	31

Fit Statistics	
R-Square	0.02398
Root MSE	145.21
Denominator DF	16

Figure: 10 The Summery Information Table of PROC SURVEYREG (Partial Table)

Tests of Model Effects			
Effect	Num DF	F Value	Pr > F
Model	2	2.04	0.1631
Intercept	1	0.98	0.3366
BMXBMI	1	1.55	0.2311
DIQ010	1	3.89	0.0662

Note: The denominator degrees of freedom for the F tests is 16.

Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-60.802376	52.4952487	-1.16	0.2638
BMXBMI	2.902695	2.3315106	1.24	0.2311
DIQ010 1	23.538083	11.9403950	1.97	0.0662
DIQ010 2	0.000000	0.0000000	.	.

Figure: 11 PROC SURVEYREG Regression Coefficient

The table (FIG:10) above listed the design of the survey. The denominator degree of freedom for F and t test is 16 due to stratification. Figure 11 showed the regression coefficient for the survey data variable DID060 (How Long Taking Insulin). The model effect table showed both the BMXBMI (Body Mass Index) and DIQ010 (Doctor Told You Have Diabetes) did not have any significant effect on the variable DID060. The regression coefficient estimates with their corresponding standard error were listed in the Estimated Regression Coefficient Table.

Conclusion:

Complex survey data works as a great source of information of any national survey. Analyzing those data gives an overall picture of population from different angles. The SAS procedures used in this paper are some of the basic steps students learn in the class room environment. In addition, SAS has more advanced procedures that can be applied to analyze any survey data very thoroughly and accurately. This paper only focused on the first few steps that can be applied for analyzing the complex survey data for any basic level SAS user.

References:

Analysing Complex Survey Data: Clustering, Stratification and Weights. (n.d.). Retrieved July 25, 2015, from http://sru.soc.surrey.ac.uk/SRU43.pdf

Lee, E. S., & Forthofer, R. N. (2006). *Analyzing Complex Survey Data*. SAGE Publications. Retrieved from https://books.google.com/books?hl=en&lr=&id=jdy_m8GWj_MC&pgis=1

Merging SAS Data Sets: Match-Merging :: Step-by-Step Programming with Base SAS(R) Software. (n.d.). Retrieved July 26, 2015, from http://support.sas.com/documentation/cdl/en/basess/58133/HTML/default/viewer.htm#a001318494.htm

SAS/STAT 9.2 User's Guide: The SURVEYMEANS Procedure (Book Excerpt) - statugsurveymeans.pdf. (n.d.). Retrieved July 26, 2015, from http://support.sas.com/documentation/cdl/en/statugsurveymeans/61837/PDF/default/statugsurveymeans.pdf

Siller, A., & Tompkins, L. (n.d.). 172-31: The Big Four: Analyzing Complex Sample Survey Data Using SAS®, SPSS, STATA, and SUDAAN - Big4.pdf. Retrieved July 25, 2015, from http://bettycjung.net/Pdfs/Big4.pdf

The SURVEYFREQ Procedure: Getting Started :: SAS/STAT(R) 9.2 User's Guide, Second Edition. (n.d.). Retrieved July 26, 2015, from http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_surveyfreq_sect002.htm

The SURVEYREG Procedure: Overview :: SAS/STAT(R) 9.22 User's Guide. (n.d.). Retrieved July 26, 2015, from http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_surveyreg_sect001.htm

Vital and Health Statistics Report Series 1, Number 56 August 2013 - sr01_056.pdf. (n.d.). Retrieved July 25, 2015, from http://www.cdc.gov/nchs/data/series/sr_01/sr01_056.pdf

CONTACT INFORMATION:

Your comment and questions are valuable and encouraged. Contact the author at:

Nushrat Alam

Masters of Public Health (Student)

Institute of Public Health

College of Pharmacy and Pharmaceutical Science

Florida Agricultural and Mechanical University

Tallahassee, Florida

Email: nushrat1.alam@famu.edu