SAS® GLOBAL FORUM 2021

Paper 1086-2021

Survey data analysis made easy with SAS®

Melanie Dove, University of California, Davis; Katherine Heck, University of California, San Francisco

ABSTRACT

Population-based, representative surveys often incorporate complex methods in data collection, such as oversampling, weighting, stratification, or clustering. If survey procedures are not used in data analyses, results will provide incorrect estimates and may overstate the results of significance testing. SAS procedures, such as PROC SURVEYFREQ and PROC SURVEYMEANS, make it easy to adjust for the complex sample design and weighting of representative surveys to obtain the correct percentages, confidence intervals, means, odds ratios, and other statistics from complex survey data.

In this introductory presentation, attendees will learn why it is necessary to use survey procedures when analyzing stratified or cluster sample surveys, what key survey design features are incorporated in the SAS code, and how to generate estimates using SAS survey procedures. This presentation will provide sample survey procedure code, and explain how to interpret the output from each survey procedure. Using examples from two publicly available surveys with different design elements (the National Health and Nutrition Examination Survey and the California Health Interview Survey), this presentation will demonstrate the following SAS survey procedures: PROC SURVEYFREQ, PROC SURVEYMEANS, PROC SURVEYLOGISTIC, and PROC SURVEYREG.

INTRODUCTION

This paper will describe key survey design features, including sampling and weighting, and illustrate how to use SAS survey procedures to adjust for these features. Two examples of publicly available health surveys will be used in this paper, the National Health and Nutrition Examination Survey (NHANES) and the California Health Interview Survey (CHIS). Briefly, NHANES is a national in-person survey and CHIS is California-based telephone survey. After discussing key survey design features, sample SAS code and output will be reviewed.

KEY SURVEY DESIGN FEATURES

A survey is a sample of individuals that is selected to represent a population. Often, there are not enough resources to collect information on the entire population of interest, so a sample of the population is selected to represent the overall population. Survey designs may incorporate information about how participants in the sample were selected, including stratification or clustering, as well as including weights that help to make the sample representative of the target population. SAS survey procedures are used to adjust for these survey design features in order to provide accurate results.

SAMPLING

Sampling indicates the method used to select participants from a target population. For example, NHANES selects approximately 5,000 participants per year from the entire US

population and CHIS selects about 26,000 participants per year from the California population to be in the survey. This paper will discuss three different types of sampling: simple random sampling, stratified sampling, and cluster sampling.

Simple Random Sampling

With this type of sample, participants are selected at random from the target population and are considered independent from each other. No adjustment in SAS is needed for this type of sampling; analyses may use PROC FREQ and other non-survey procedures.

Stratified Sampling

With stratified sampling, the target population is divided into different groups (or strata) and participants are sampled from each group. This ensures that there are enough people in the sample from each group. For example, in CHIS, the target population (California) is divided into different geographic regions (or strata) and participants are randomly selected from each region. This ensures that there are participants from all areas of California in the sample.

With stratified sampling, participants are not independent from each other, and the STRATA statement is used to adjust for the fact that individuals within strata are more similar to each other compared with individuals from different strata. Not adjusting for this fact will lead to an overestimate of the variance and significance of findings.

Cluster Sampling

With cluster sampling, the target population is also divided into different groups, based on a shared characteristic, such as the school they attend or the county in which they live. The group itself (the cluster) is selected randomly, and then participants are drawn from that cluster. For example, in NHANES, counties are selected to be in the survey from a list of counties in the target population (United States), and participants are selected from each county or cluster.

Similar to stratified sampling, participants are not independent of each other. The CLUSTER statement is used to adjust for the fact that individuals within clusters are more similar to each other compared with individuals from different clusters.

WEIGHTING

A weight is a value indicating the number of people the respondent represents. Each person is given one or more weights so that the weighted estimates are representative of the target population. The SAS statement WEIGHT corrects for the following factors:

- Differing probability of sampling different strata or clusters. Participants may be more likely to be included in the survey based on the design of the survey than would otherwise with simple random sampling. For example, participants from rural geographic regions may be more likely to be included in the sample if it is stratified by geography compared with simple random sampling.
- Nonresponse – People who respond to the survey may be different than people who do not respond to the survey. For example, older adults who are retired may have more spare time and be more likely to respond to the survey, compared with younger adults.
- Noncoverage – The sampling frame may not include the entire population. For example, for a survey done by calling cell phone numbers, people without cell phones would not be included in the sampling frame. Some surveys adjust for this error.
- Some surveys also use raking of weights to more closely align weighted data to population totals for characteristics such as age or region.

This paper will describe two methods of creating weights: single weights and replicate weights.

Single Weights

With single weights there is one weight per person. The weight represents the number of individuals in the population that the sampled person represents.

Replicate Weights

With replicate weights, there is more than one weight per person. Often, each participant receives a base weight and an additional 80 replicate weights. Replicate weights may be used when there is a concern about confidentiality, since releasing a stratification variable may reveal information about participants. For example, in CHIS the target population is divided into different strata based on geography. Instead of releasing a strata variable with information on each participant's geographic location, they use replicate weights to account for both the stratification and weight. With replicate weights, two SAS statements are used: WEIGHT to list the base weight and REPWEIGHT to list the replicate weights.

SURVEY STATEMENTS

All SAS survey procedures use common statements to identify survey design and weighting elements.

- PROC statement: This statement identifies the survey procedure being used (i.e. PROC SURVEYFREQ or PROC SURVEYLOGISIC), and includes options such as identifying the data set, how to handle missing values, the finite population correction if included in the data set, and the variance estimation method, such as jackknife or Taylor series linearization.
- CLUSTER statement: Identifies the clustering variable(s).
- STRATA statement: identifies the stratification variable(s).
- WEIGHTS and REPWEIGHT statements: identify the weighting variables; options include specifying coefficient values for jackknife variance estimation or replicate weights.
- DOMAIN: in survey procedures, used instead of the WHERE or BY statement to provide analyses of subgroups, such as performing analyses separately for men and women. If used in survey procedures to subset a population, the BY statement will provide inaccurate estimates of the standard errors and confidence limits; DOMAIN allows for accurate subgroup analyses.

Additional statements are similar to those used in non-survey SAS procedures, such as TABLES, CLASS, CONTRAST, VAR, MODEL, and TEST; see SAS online documentation for options available in each procedure.

EXAMPLES

To demonstrate how to analyze survey data with SAS, this paper will use examples from NHANES and CHIS, which have different sampling and weighting methods, as outlined in Table 1 below.

Survey	Sampling method	Weight method	SAS statements
NHANES	Stratification and clustering	Single weights	STRATA CLUSTER WEIGHT
CHIS	Stratification	Replicate weights	WEIGHT REPWEIGHT

Table 1. Example Surveys, Sampling and Weight Methods, and SAS Statements

Below are two general syntax examples for how to adjust for the survey design features, using NHANES and CHIS data:

- NHANES data:

```
proc surveyfreq data=datasetname varmethod=taylor;
  strata stratavariable;
  cluster clustervariable;
  weight weightvariable;
    tables variablename;
run;
```

- CHIS data:

```
proc surveyfreq data=datasetname varmethod=jackknife;
  weight baseweightvariable;
  repweight replicateweightvariables / jkcoefs=1;
    tables variablename;
run;
```

Below are more specific examples of SAS code and output for four different survey procedures: 1) PROC SURVEYFREQ, 2) PROC SURVEYMEANS, 3) PROC SURVEYLOGISTIC, and 4) PROC SURVEYREG. The strata, cluster, and weight variables below may change when using different years of data. Make sure to consult each survey's documentation for the correct variables to include in the code.

PROC SURVEYFREQ

In the following example, CHIS data is used to estimate the percent of men and women ('srsex') with hypertension ('ab29'). The DOMAIN statement is not available with the SURVEYFREQ procedure. Therefore, the gender variable is included as the first variable in the cross tab to get estimates by gender. Do not subset the data using a WHERE or BY statement. The following options are specified on the tables statement:

- CL to get 95% confidence limits
- ROW to get row percentages (can also use COL for column percentages)
- NOTOTAL suppresses the total estimates, and
- CHISQ requests the chi-square statistic and p-value

```
proc surveyfreq data=CHIS varmethod=jackknife;
  weight rakedw0;
  repweight rakedw1-rakedw80 / jkcoefs=1;
    tables srsex*ab29 / cl row nototal chisq;
run;
```

Data Summary	
Number of Observations	21055
Sum of Weights	29390199.7

Variance Estimation	
Method	Jackknife
Replicate Weights	ADULT
Number of Replicates	80

Table of SRSEX by AB29

SRSEX	AB29	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	95% Confidence Limits for Percent		Row Percent	Std Err of Row Percent	95% Confidence Limits for Row Percent	
Male	Yes	3747	4457319	174688	15.1660	0.5944	13.9832	16.3488	31.0260	1.2159	28.6061	33.4458
	No	5420	9759533	169867	33.2068	0.5780	32.0566	34.3570	67.9329	1.1824	65.5799	70.2860
	Borderline HTN	140	149569	33823	0.5089	0.1151	0.2799	0.7379	1.0411	0.2354	0.5726	1.5096
Female	Yes	4395	3878467	127700	13.1965	0.4345	12.3318	14.0611	25.8155	0.8500	24.1240	27.5071
	No	7245	11053943	127459	37.6110	0.4337	36.7479	38.4740	73.5763	0.8484	71.8880	75.2647
	Borderline HTN	108	91368	21963	0.3109	0.0747	0.1622	0.4596	0.6082	0.1462	0.3172	0.8991

Rao-Scott Chi-Square Test	
Pearson Chi-Square	86.0484
Design Correction	5.8150
Rao-Scott Chi-Square	14.7978
DF	2
Pr > ChiSq	0.0006
F Value	7.3989
Num DF	2
Den DF	160
Pr > F	0.0008
Sample Size = 21055	

Output 1. Output from a PROC SURVEYFREQ Statement

From Output 1, 31.0% of men and 25.8% of men had hypertension. These estimates are statistically different as indicated by the chi-square p-value of 0.0006.

PROC SURVEYMEANS

The following example requests the mean number of times walked for leisure in the past week ('ad41W') for different age categories ('srage_p1'), also using CHIS data. The continuous variable 'number of times walked' is included on the VAR statement. The DOMAIN statement is used for the categorical age variable.

```
proc surveymeans data=CHIS varmethod=jackknife;
  weight rakedw0;
  repweight rakedw1-rakedw80 / jkcoefs=1;
    var ad41W;
    domain SRAGE_P1;
run;
```

The SAS System

The SURVEYMEANS Procedure

Statistics for SRAGE_P1 Domains							
SRAGE_P1	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
18-29	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	2802	3.016137	0.213305	2.59164564	3.44062809
30-49	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	4587	2.687602	0.152270	2.38457612	2.99062815
50-69	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	8337	2.900668	0.149342	2.60346776	3.19786896
70+	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	5329	2.534561	0.205938	2.12473082	2.94439086

Output 2. Output from a PROC SURVEYMEANS Statement

From Output 2, the mean number of times walked for leisure in the past seven days was 3.0 times for 18-29 year-olds and 2.5 times for adults 70 years and older.

PROC SURVEYLOGISTIC

The following example examines the association between insurance status ('uninsured') and not having a usual source of healthcare ('nousual'). The CLASS statement is used for the categorical variable 'uninsured' and the 'Insured' category is specified as the referent. On the model statement, the '(descending)' option is used so that SAS estimates the odds that 'nousual' equals 1 (or does not have a usual source of healthcare) instead of 0 (has a usual source of healthcare).

```
proc surveylogistic data=CHIS varmethod=jackknife;
  weight rakedw0;
  cluster sdmvpsu;
  repweight rakedw1-rakedw80 / jkcoefs=1;
    class uninsured (ref='Insured')/ param=ref;
    model nousual (descending)=uninsured;
run;
```

Class Level Information		
Class	Value	Design Variables
uninsured	Insured	-1
	Uninsured	1

Variance Estimation	
Method	Jackknife
Replicate Weights	ADULT
Number of Replicates	80

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	22890096	21137570
SC	22890104	21137585
-2 Log L	22890094	21137566

Testing Global Null Hypothesis: BETA=0				
Test	F Value	Num DF	Den DF	Pr > F
Likelihood Ratio	1752529	1	Infty	<.0001
Score	57.68	1	80	<.0001
Wald	89.21	1	80	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
uninsured	1	89.2128	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-1.0612	0.0759	-13.98	<.0001
uninsured	Uninsured	0.8232	0.0872	9.45	<.0001
NOTE: The degrees of freedom for the t tests is 80.					

Odds Ratio Estimates			
Effect	Point Estimate	95% Confidence Limits	
uninsured Uninsured vs Insured	5.188	3.668	7.340
NOTE: The degrees of freedom in computing the confidence limits is 80.			

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	33.7	Somers' D	0.275
Percent Discordant	6.3	Gamma	0.686
Percent Tied	60.0	Tau-a	0.071
Pairs	22790240	c	0.637

Output 3. Output from a PROC SURVEYLOGISTIC Statement

From Output 3, the odds ratio is 5.2 with a 95% confidence interval of 3.7 to 7.3, suggesting that uninsured adults are 5 times as likely to not have a usual source of healthcare, compared with insured adults.

PROC SURVEYREG

Using the NHANES data, the following example examines the association between health insurance status ('hi') and cotinine ('lbxcot'), a biomarker of nicotine measured in ng/mL. The DOMAIN statement is used with the variable 'set', a flag variable that equals 1 if in the analysis group and 0 otherwise. On the model statement, the option SOLUTION is included to get the beta estimates and the option CLPARM is used to get the 95% confidence intervals.

```
proc surveyreg data=NHANES varmethod=taylor;
  strata sdmvstra;
  cluster sdmvpsu;
  weight wtint2yr;
    domain set;
    class hi;
    model lbxcot=hi / solution clparm;
run;
```

The SAS System

The SURVEYREG Procedure

set=1

Domain Regression Analysis for Variable LBXCOT

Domain Summary	
Number of Observations	19984
Number of Observations in Domain	19984
Number of Observations Not in Domain	0
Sum of Weights in Domain	173040335
Weighted Mean of LBXCOT	65.85795
Weighted Sum of LBXCOT	1.13961E10

Fit Statistics	
R-Square	0.02606
Root MSE	131.37
Denominator DF	79

Tests of Model Effects			
Effect	Num DF	F Value	Pr > F
Model	3	68.17	<.0001
Intercept	1	715.99	<.0001
hi	3	68.17	<.0001

Note: The denominator degrees of freedom for the F tests is 79.

Estimated Regression Coefficients						
Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Interval	
Intercept	88.954231	4.13942163	21.49	<.0001	80.714918	97.193544
hi medicaid	28.249178	5.49630214	5.14	<.0001	17.309062	39.189294
hi other	-2.501724	5.72169621	-0.44	0.6631	-13.890476	8.887027
hi private	-38.584854	3.85904640	-10.00	<.0001	-46.266094	-30.903615
hi uninsured	0.000000	0.00000000	.	.	0.000000	0.000000

Output 4. Output from a PROC SURVEYREG Statement

From Output 4, the average cotinine level for adults with Medicaid insurance is 28.2 ng/mL higher than uninsured adults. The average cotinine level for adults with private insurance is 38.6 ng/mL lower than uninsured adults.

CONCLUSION

The SAS survey procedures provide a flexible way to obtain accurate results of analyses using stratified or clustered survey data.

REFERENCES

Centers for Disease Control and Prevention, National Center for Health Statistics. "NHANES Questionnaires, Datasets, and Related Documentation." Accessed April 14, 2021. https://wwwn.cdc.gov/nchs/nhanes/Default.aspx.

UCLA Center for Health Policy Research, California Health Interview Survey. "Public Use Data." Accessed April 14, 2021. https://healthpolicy.ucla.edu/chis/data/Pages/GetCHISData.aspx

CONTACT INFORMATION

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

Melanie Dove
University of California, Davis
mdove@ucdavis.edu

Katherine Heck
University of California, San Francisco
katherine.heck@ucsf.edu