Paper 4635-2020

Survey Data Analysis Made Easy with SAS®

Melanie Dove, UC Davis; Katherine Heck, UC San Francisco

ABSTRACT

Population-based, representative surveys often incorporate complex methods in data collection, such as oversampling, weighting, stratification, or clustering. Analysis of these data sets using standard procedures (such as the FREQ procedure) results in incorrect estimates and might overstate the statistical significance of results due to the complex survey design factors. However, SAS® survey procedures, such as the SURVEYFREQ and SURVEYMEANS procedures, make it easy to adjust for the complex sample design and weighting of representative surveys. This hands-on workshop (HOW) provides an overview of complex survey design and explains how SAS survey procedures can adjust for complex survey design factors. Attendees learn how to easily generate accurate frequencies, percentages, means, and odds ratios from survey data sets using SAS survey procedures. The workshop provides information about obtaining accurate standard errors and confidence intervals, and demonstrates how to statistically test for differences using chi-square or t-tests. The course also explains how to interpret the output data from the survey procedures and provides examples of SAS code and output. This workshop uses publicly available data from the National Health and Nutrition Examination Survey (NHANES) and the California Health Interview Survey (CHIS) as examples. Attendees have the opportunity to practice using SAS survey procedures on these data sets.

INTRODUCTION

This paper describes four SAS® procedures to analyze survey data, SURVEYFREQ, SURVEYMEANS, SURVEYLOGISTIC, and SURVEYREG, with examples using data from the California Health Interview Survey (CHIS). The first procedure described, PROC SURVEYFREQ, includes the most detail about how to adjust for the survey design factors, and the rest of the procedures use the same set of code to adjust for these factors. The audience will gain skills in understanding the design of complex sample surveys, and the analysis of survey data sets using SAS survey procedures.

WHY WE USE SURVEY PROCEDURES

Random sampling results in a sample that is representative of a population, within a margin of error. However, random samples may not result in large enough numbers for accurate estimates of smaller subpopulations, and may be cost prohibitive. Cluster sampling or stratification methods are used to sample respondents from different subgroups – for example, people who live in different counties or attend different schools – at varying rates, enabling data collection of adequate sample sizes for smaller subgroups. In cluster sampling, respondents are selected from a 'cluster' such as a school or household. In stratified sampling, specified numbers of respondents are selected from strata that are created based on characteristics, such as county. For example, a stratified sample might select 200 people from Los Angeles County and 100 people from San Francisco County.

Stratification and cluster sampling methods can result in a smaller sample that becomes representative of the target population when weights are applied. People who were sampled at lower rates receive higher weights to make the sample representative when weighted. However, because sampling probabilities varied between different clusters or strata, survey procedures are needed to correctly calculate the variance. If survey methods are not used when analyzing stratified or cluster data, standard errors, confidence intervals, and significance levels of statistics will be incorrect.

EXAMPLE DATA

To demonstrate the concepts in this paper, we use data from the 2018 California Health Interview Survey (CHIS), a representative sample of California's non-institutionalized population. CHIS is a telephone survey that began collecting data every other year in 2001 and every year in 2011. CHIS is conducted by the University of California Los Angeles Center for Health Policy Research, and is the largest state-level health survey in the United States. Each year three data sets are available, one for adults, teens, and children, all of which are available for download on the CHIS website. In our examples, we only use the adult (>=18 years) data set.

CHIS uses a two-stage geographically stratified random-digit-dial sample design. In the first stage, telephone numbers are randomly sampled within counties. In the second stage, individuals are sampled from each household. For their publicly available data sets, CHIS provides replicate weights, which are a series of weight variables that must be used in combination to correctly weight the sample. The final weight variable (rakedw0) ensures that estimates are representative of the California population and the replicate weights (rakedw1 – rakedw80) ensure that the variance is correctly estimated. Replicate weights are used in place of the geographic stratification variable because of confidentiality concerns in releasing county-level data. However, the stratification variable is available in the confidential CHIS data, which can be accessed through their Data Access Center. In the first example (PROC SURVEYFREQ), we provide code for how to analyze both the confidential and public CHIS data.

In the following examples, we use the categorical variables of current use of e-cigarettes (yes, no) and age (18-25, 26-29, 30-34, and 35+), and the continuous variable BMI. To prepare the data, we completed the following steps:

- Created a new data set called 'chis' where we kept only the variables that we needed for the analysis.
- Created a new variable called 'ecig_curr' that combines ever (ac81c) and current (ac82c_p1) e-cigarette use. Current (past 30 days) e-cigarette users are classified as '1' and non-users as '0', which is how we want this variable categorized for the examples.
- Created a new variable called 'age' with four categories – 18-25, 26-29, 30-34, and 35+.
- Created a new variable called 'bmi' that sets body mass index (BMI) values over 100 to missing.

We used the following SAS code to create the data set:

```
proc format;

    value agef  1='18-25'
                2='26-29'
                3='30-34'
                4='35+';
run;

data chis (keep = ac81c ac82c_p1 ecig_curr rakedw0 rakedw1-rakedw80
srage_p1 age BMI_P bmi);
set chis.adult;

    /*create ecig_curr variable*/
```

```
    if ac81c= 1 then do;
       if ac82c_p1 in (2,3,4,5) then ecig_curr=1;
       else if ac82c_p1 =1 then ecig_curr=0;
    end;
    else if ac81c=2 then ecig_curr=0;

   /*create categorical age variable*/
    if srage_p1=18 then age=1;
    else if srage_p1=26 then age=2;
    else if srage_p1=30 then age=3;
    else age=4;

    /*set outliers from BMI to missing*/
    if BMI_P >100 then bmi=.;
    else bmi=BMI_P;

format age agef.;
run;
```

PROC SURVEYFREQ

The SURVEYFREQ procedure is used to output frequency tables, percentages, confidence intervals, and test statistics such as chi-square, using stratified or clustered survey data. PROC SURVEYFREQ is similar to the FREQ procedure, but includes statements to specify the survey-related variables, such as stratum, cluster, weight and/or replicate weights (repweight), and the variance estimation method. Whether or not to include these options depends on the design of the survey. Surveys often provide documentation that describes their design and sample code for how to analyze their data (resources for CHIS are provided in the references).

For this first procedure, we provide code used to analyze both the confidential and publicly available CHIS data in order to demonstrate several survey design features that are not available in the public data, including the STRATA and CLUSTER statements. For the rest of the procedures, we only provide sample code for the public data set.

The first set of SAS code below demonstrates how to analyze the confidential CHIS data, which uses the Taylor series method to calculate the variance. The variance method is specified in the first line of code (VARMETHOD=TAYLOR), along with the option 'NOMCAR' to specify the assumption that missing values are not completely at random. The next several lines of code include a strata variable (STRATA tsvarstr) to account for the geographic stratification sample design, a cluster variable (CLUSTER tsvrunit) to account for the fact that people living in a household are clustered (only used if combining the children, teen, and adult data), and one weight variable (WEIGHT rakedw0). The TABLES statement requests the frequency and percent of current e-cigarette use by age. The options in the TABLES statement request row percentages (row) and 95% confidence intervals (cl) for the row percentages.

```
PROC SURVEYFREQ DATA=chis NOMCAR VARMETHOD=TAYLOR;
   STRATA tsvarstr;
   CLUSTER tsvrunit;
   WEIGHT rakedw0;
   TABLE age*ecig_curr/row cl;
RUN;
```

The next example uses the publicly available CHIS data to examine the frequency of current e-cigarette use by age. Because the strata variable isn't provided in the public use data sets (due to confidentiality around releasing geographic variables), CHIS uses replicate weights with jackknife variance estimation. The code below specifies the variance estimation method (VARMETHOD=JACKKNIFE), final weight variable (WEIGHT rakedw0), and replicate weights (REPWEIGHT rakedw1-rakedw80). The 'JKCOEFS=1' option is necessary to obtain accurate variance estimates. A 'CHISQ' option is added to the 'TABLES' statement to test whether there is an association between age and e-cigarette use.

```
PROC SURVEY FREQ DATA=chis VARMETHOD=JACKKNIFE;
    WEIGHT rakedw0;
    REPWEIGHT rakedw1-rakedw80/ JKCOEFS=1;
    TABLES age*ecig_curr/ROW CL CHISQ;
RUN;
```

The results are shown below and include the variables in the TABLES statement on the left, followed by the unweighted frequency and the weighted frequency. In the second data row, the 300 e-cigarette users aged 18-25 when weighted represent a population size of 682,369. Also included in the table is the standard error of the weighted frequency (65,389), and the overall weighted percent in that cell (2.3%), with standard error and confidence limits for that value. The row percent, standard error, and 95% confidence limits are shown next. Using the row percentages, among adults aged 18-25, 14.9% (95% CI: 12.1%, 17.7%) currently use e-cigarettes, whereas only 2.6% (95% CI: 2.1%, 3.1%) of adults 35 years or older use e-cigarettes. From the second table in the output, the 'Pr>ChiSq', is <0.0001, indicating that there is a statistically significant association between age and e-cigarette use.

Table of age by ecig_curr

age	ecig_curr	Frequency	Weighted Frequency	Std Err 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	
18-25	0	1723	3895203	73499	13.1158	0.2475	12.6233	13.6083	85.0932	1.3941	82.3189	87.8676
	1	300	682369	65389	2.2976	0.2202	1.8595	2.7358	14.9068	1.3941	12.1324	17.6811
	Total	2023	4577572	54089	15.4134	0.1821	15.0510	15.7759	100.000			
26-29	0	713	1770416	69511	5.9613	0.2341	5.4955	6.4271	86.7837	2.1934	82.4186	91.1488
	1	109	269617	44256	0.9078	0.1490	0.6113	1.2044	13.2163	2.1934	8.8512	17.5814
	Total	822	2040033	54089	6.8691	0.1821	6.5067	7.2316	100.000			
30-34	0	911	2315617	84070	7.7971	0.2831	7.2337	8.3604	89.6822	1.9870	85.7281	93.6364
	1	103	266407	50758	0.8970	0.1709	0.5569	1.2372	10.3178	1.9870	6.3636	14.2719
	Total	1014	2582024	66486	8.6941	0.2239	8.2486	9.1396	100.000			
35+	0	16905	19961171	84709	67.2125	0.2852	66.6448	67.7801	97.3764	0.2525	96.8739	97.8788
	1	413	537819	51709	1.8109	0.1741	1.4644	2.1574	2.6236	0.2525	2.1212	3.1261
	Total	17318	20498990	66486	69.0234	0.2239	68.5779	69.4689	100.000			
Total	0	20252	27942407	123992	94.0866	0.4175	93.2557	94.9174				
	1	925	1756212	123992	5.9134	0.4175	5.0826	6.7443				
	Total	21177	29698619	2.72926E-6	100.000							

Rao-Scott Chi-Square Test	
Pearson Chi-Square	962.4680
Design Correction	5.3229
Rao-Scott Chi-Square	180.8160
DF	3
Pr > ChiSq	<.0001
F Value	60.2720
Num DF	3
Den DF	240
Pr > F	<.0001
Sample Size = 21177	

Output 1. Output from PROC SURVEYFREQ

When using survey procedures, formats or 'flag' variables should be used to identify a group of interest for analysis, rather than using a WHERE, IF, or BY statement to subset a sample. Standard errors will be incorrect if the survey procedure code does not include the whole sample. For example, to find the percent of e-cigarettes users aged 18-25, we would NOT want to use a statement such as 'WHERE age=1'.

PROC SURVEYMEANS

The SURVEYMEANS procedure is similar to PROC SURVEYFREQ in its structure, but produces means, medians, and other statistics for continuous or categorical variables. By default, the sample size, mean, standard error, and 95% confidence interval are included in the output. Other statistics such as medians, percentiles, or t-tests can be requested in the PROC SURVEYMEANS statement. The VAR statement identifies the variable of interest, and the DOMAIN statement requests analysis for subpopulations in addition to the entire study population. By including a variable on the DOMAIN statement, SAS will output a table with the number of respondents in each DOMAIN category and their mean values. In the example below, we request the mean BMI for each category of e-cigarette use.

```
PROC SURVEYMEANS DATA=chis VARMETHOD=JACKKNIFE;
    WEIGHT rakedw0;
    REPWEIGHT rakedw1-rakedw80/ JKCOEFS=1;
    VAR bmi;
    DOMAIN ecig_curr;
RUN;
```

Statistics					
Variable	N	Mean	Std Error of Mean	95% CL for Mean	
bmi	21175	27.485971	0.096833	27.2932680	27.6786745

Domain Statistics in ecig_curr						
ecig_curr	Variable	N	Mean	Std Error of Mean	95% CL for Mean	
0	bmi	20250	27.526904	0.096549	27.3347647	27.7190431
1	bmi	925	26.834879	0.389685	26.0593806	27.6103774

Output 2. Output from PROC SURVEYMEANS

From the output above, we see that the mean BMI for the entire sample is 27.5. Because we included the variable for e-cigarette use on the DOMAIN statement, we also get the mean BMI for e-cigarette users (26.8) and non-users (27.5).

PROC SURVEYMEANS can also be used for categorical variables with the use of the CLASS statement. In the SAS code below we request the mean of adults who currently use e-cigarettes; since non-smokers are coded to 0 and smokers to 1, the mean value is equivalent to a percentage.

```
PROC SURVEYMEANS DATA=chis VARMETHOD=JACKKNIFE;
    WEIGHT rakedw0;
    REPWEIGHT rakedw1-rakedw80/ JKCOEFS=1;
    CLASS ecig_curr;
    VAR ecig_curr;
RUN;
```

PROC SURVEYLOGISTIC

To examine whether there is an association between age, BMI, and current e-cigarette use, the SURVEYLOGISTIC procedure can be used. Similar code is used to specify the variance method, final weight (rakedw0), and replicate weights (rakedw1-rakedw80) as used in PROC SURVEYFREQ and PROC SURVEYMEANS. The dependent variable (ecig_curr) is a 0/1 variable and we specify the option 'descending' on the model statement so that SAS models the probability of being a current e-cigarette user (ecig_curr=1) rather than the default (ecig_curr=0). We include the variable 'age' on the class statement so that SAS treats age as a categorical variable. We do not include BMI on the class statement and SAS will consider BMI as a continuous variable.

```
PROC SURVEYLOGISTIC DATA=chis VARMETHOD=JACKKNIFE;
    WEIGHT rakedw0;
    REPWEIGHT rakedw1-rakedw80/ JKCOEFS=1;
    CLASS age;
    MODEL ecig_curr (descending)=age bmi;
RUN;
```

Below we include the output for the odds ratio and 95% confidence intervals:

Odds Ratio Estimates			
Effect	Point Estimate	95% Confidence Limits	
age 18-25 vs 35+	6.469	4.892	8.553
age 26-29 vs 35+	5.644	3.620	8.798
age 30-34 vs 35+	4.269	2.751	6.624
bmi	0.998	0.979	1.017

Output 3. Output from PROC SURVEYLOGISTIC

Adults aged 18-25 years have 6.5 (95% CI: 4.9, 8.6) times the odds of currently using e-cigarettes, compared with adults 35 years and older, controlling for BMI. Adults aged 26 to 34 also have a significantly increased odds compared with adults 35 years and older. There is no association between BMI and e-cigarette use because the 95% confidence interval includes 1 (OR=0.998, 95% CI: 0.98, 1.02).

SAS uses the highest value of an independent variable as the referent category. You can change the referent category by including the "ref= / param=ref" option in the CLASS statement.

```
PROC SURVEYLOGISTIC DATA=chis VARMETHOD=JACKKNIFE;
    WEIGHT rakedw0;
    REPWEIGHT rakedw1-rakedw80/ JKCOEFS=1;
    CLASS age (ref="18-25")/ param=ref;
    MODEL ecig_curr (descending)=age bmi;
RUN;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Confidence Limits	
age 26-29 vs 18-25	0.872	0.562	1.354
age 30-34 vs 18-25	0.660	0.435	1.001
age 35+ vs 18-25	0.155	0.117	0.204
bmi	0.998	0.979	1.017
NOTE: The degrees of freedom in computing the confidence limits is 80.			

Output 4. Output from PROC SURVEYLOGISTIC Using a Different Reference Category

PROC SURVEYREG

You can use the SURVEYREG procedure to examine whether mean BMI is different for e-cigarette users and non-users, adjusting for age. You can change the referent category of the variables on the CLASS statement using the same method as in the PROC SURVEYLOGISTIC example. The '/SOLUTION' option at the end of the MODEL statement tells SAS to include the regression coefficients in the output. The CLPARM options tells SAS to include the 95% confidence intervals of the parameter estimates in the output.

```
PROC SURVEYREG DATA=chis VARMETHOD=JACKKNIFE;
    WEIGHT rakedw0;
    REPWEIGHT rakedw1-rakedw80/ JKCOEFS=1;
    CLASS ecig_curr age;
    MODEL bmi=ecig_curr age/ SOLUTION CLPARM;
RUN;
```

Estimated Regression Coefficients						
Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Interval	
Intercept	27.7799923	0.39367380	70.57	<.0001	26.9965564	28.5634281
ecig_curr 0	0.0889573	0.38865617	0.23	0.8195	-0.6844932	0.8624077
ecig_curr 1	0.0000000	0.00000000	.	.	0.0000000	0.0000000
age 18-25	-2.2375279	0.24472852	-9.14	<.0001	-2.7245532	-1.7505026
age 26-29	-0.5524403	0.44441139	-1.24	0.2175	-1.4368471	0.3319666
age 30-34	0.0598630	0.54648565	0.11	0.9130	-1.0276781	1.1474041
age 35+	0.0000000	0.00000000	.	.	0.0000000	0.0000000

Output 5. Output from PROC SURVEYREG

Adults who currently do not use e-cigarettes have BMI values that are on average 0.089 kg/m^2 higher than adults who currently use e-cigarettes, controlling for age. This difference is not statistically significant (p=0.8195). Adults age 18-25 have BMI values that are on average 2.24 kg/m^2 lower than adults aged 35 years or older. The 18-25 year age group is the only age group with a mean BMI statistically different than the referent of 35+ (p=<0.0001).

CONCLUSION

With the use of SAS survey procedures, anyone can analyze survey data. Survey design factors can be adjusted for in a variety of procedures, including SURVEYFREQ, SURVEYMEANS, SURVEYLOGISTIC, and SURVEYREG. Adjusting for the survey design factors in SAS will produce representative estimates and correct standard errors.

REFERENCES

UCLA Center for Health Policy Research. "Get CHIS Data". Accessed February 26, 2020. http://healthpolicy.ucla.edu/chis/data/Pages/GetCHISData.aspx

UCLA Center for Health Policy Research. "CHIS Methodology Documentation". Accessed February 26, 2020. http://healthpolicy.ucla.edu/chis/design/Pages/methodology.aspx

UCLA Center for Health Policy Research. "Analyze CHIS Data – Sample Code". Accessed February 26, 2020. https://healthpolicy.ucla.edu/chis/analyze/Pages/sample-code.aspx

CONTACT INFORMATION

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

Melanie Dove, ScD, MPH
University of California, Davis
Department of Public Health Sciences
Division of Health Policy and Management
530-754-0912
mdove@ucdavis.edu