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:

```sas
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.
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.

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

**Output 1. Output from PROC SURVEYFREQ**

**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.

```sas
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.

```sas
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](image)

**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;

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

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;

Output 5. Output from PROC SURVEYREG

Adults who currently do not use e-cigarettes have BMI values that are on average 0.089 kg/m² 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/m² 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


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