

## Statistical Analyses of Public Health Surveys Using SAS® Survey Package

Trung Ha and Julia Soulakova, Burnett School of Biomedical Sciences, College of Medicine,  
University of Central Florida

### ABSTRACT

We will discuss the key SAS® survey package (SAS 9.4 or later) procedures that one can use to perform statistical analyses of complex survey data with replicate weights. We will address the topic of replicate weights and balanced repeated replications for variance estimation. The key methods for obtaining summary statistics and performing testing for quantitative and qualitative outcomes will be discussed. In addition, model building approaches for logistic regressions with fixed effects will be addressed. Examples will illustrate the concepts and procedures for smoking-related behaviors using the Tobacco Use Supplement (TUS) to the Current Population Survey (CPS) data set with 160 replicate weights. The survey is widely used to estimate the prevalence of smoking, as well as rates of smoking-related events and behaviors, e.g., quitting behaviors among smokers and attitudes towards smoking laws.

Keywords: complex survey, balanced repeated replication, CPS, TUS

### INTRODUCTION

The Current Population Survey (CPS) has been administrated monthly by the U.S. Census Bureau since 1942. It is the primary source of labor force information in the US. The CPS has a number of Supplements, such as the Tobacco Use Supplement (TUS), which are widely used in policy-making and research. The CPS uses a monthly sample of approximately 60,000 households. This sample represents the civilian noninstitutional population (US Department of Commerce, Census Bureau, 2016). The CPS sample is drawn using a multi-stage sampling scheme (U.S. Bureau of Labor Statistics & U.S. Census Bureau, 2006, Chapter 3). Thus, statistical analyses of the CPS and Supplements require proper adjustments for complex sampling. First, in addition to the data sets, the main weight and replicate weights – made available for public use by the U.S. Census Bureau – should be downloaded. Next, the balanced repeated replication (BRR) method with a Fay correction of 0.5 and the replicate weights should be used to estimate variance (Fay & Train, 1995; Judkins, 1990; US Department of Commerce, Census Bureau, 2016; Wolter, 2007).

To illustrate the survey procedures in SAS for analyzing 2014-15 TUS-CPS data, we downloaded the three data sets and three weight sets for July 2014, January 2015, and May 2015. We then merged the six sets and rescaled the weights. The final main weight was denoted by RepWgt0, and the final 160 replicate weights were denoted by RepWgt1, RepWgt2,..., RepWgt160. We tested for the association between two categorical measures while ignoring the potential effect of the other measures, as well as testing for the association while controlling for the effect of the other measures. The first approach was based on a Rao-Scott chi-square test for a contingency table (Rao & Scott, 1981), while the second approach was based on a multiple logistic regression for survey data (Agresti, 2013).

### TESTING FOR THE ASSOCIATION BETWEEN TWO CATEGORICAL MEASURES

Suppose we want to test for the association between the smoking status (denoted by SmokStat) and sex (denoted by Sex), where the smoking status is categorized into four levels: never smokers, former smokers, occasional smokers, and daily smokers; and sex is binary (male, female). The following statements in the SURVEYFREQ procedure call the chi-square tests for association, and percentages for cross groups and overall groups:

```
PROC SURVEYFREQ data=TUS_CPS varmethod=brr(fay);  
  table SmokStat*Sex /chisq col row;  
  weight RepWgt0;  
  repwgt RepWgt1-RepWgt160;  
RUN;
```

We use VARMETHOD=brr(fay) to request the balanced repeated replication (BRR) method with a Fay correction of 0.5. Note that specifying “fay” is equivalent to specifying “fay=0.5”. We specify the replicate weights using the REPWGT statement. We use the WEIGHT statement to supply the main weight RepWgt0. The CHISQ option requests a Rao-Scott chi-square goodness-of-fit test. The ROW and COL options request frequencies and percentages by row and column, respectively. Figure 1 - Figure 4 display the output.

| Data Summary           |           |
|------------------------|-----------|
| Number of Observations | 158276    |
| Sum of Weights         | 233425826 |

Figure 1. TUS-CPS Data Summary

| Variance Estimation  |         |
|----------------------|---------|
| Method               | BRR     |
| Replicate Weights    | TUS_CPS |
| Number of Replicates | 160     |
| Fay Coefficient      | 0.500   |

Figure 2. Variance Estimation Method

| Table of SMOKSTAT by sex |        |           |                    |                     |         |                    |             |                        |                |                        |
|--------------------------|--------|-----------|--------------------|---------------------|---------|--------------------|-------------|------------------------|----------------|------------------------|
| SMOKSTAT                 | Sex    | Frequency | Weighted Frequency | Std Dev of Wgt Freq | Percent | Std Err of Percent | Row Percent | Std Err of Row Percent | Column Percent | Std Err of Col Percent |
| Never smoker             | Male   | 41993     | 70578729           | 246018              | 30.2360 | 0.0979             | 44.5726     | 0.1039                 | 63.0479        | 0.1945                 |
|                          | Female | 60646     | 87766827           | 246552              | 37.5995 | 0.0988             | 55.4274     | 0.1039                 | 72.2472        | 0.1746                 |
|                          | Total  | 102639    | 158345556          | 368801              | 67.8355 | 0.1379             | 100.000     |                        |                |                        |
| Daily smoker             | Male   | 8574      | 12898014           | 143191              | 5.5255  | 0.0609             | 52.9918     | 0.4089                 | 11.5218        | 0.1265                 |
|                          | Female | 8894      | 11441646           | 153960              | 4.9016  | 0.0660             | 47.0082     | 0.4089                 | 9.4184         | 0.1272                 |
|                          | Total  | 17468     | 24339660           | 222190              | 10.4271 | 0.0947             | 100.000     |                        |                |                        |
| Occasional smoker        | Male   | 2332      | 4000223            | 98947               | 1.7137  | 0.0424             | 56.4253     | 0.8398                 | 3.5734         | 0.0885                 |
|                          | Female | 2295      | 3089186            | 67162               | 1.3234  | 0.0288             | 43.5747     | 0.8398                 | 2.5429         | 0.0554                 |
|                          | Total  | 4627      | 7089409            | 115212              | 3.0371  | 0.0495             | 100.000     |                        |                |                        |
| Former smoker            | Male   | 17713     | 24467566           | 179237              | 10.4819 | 0.0765             | 56.0524     | 0.2465                 | 21.8569        | 0.1588                 |
|                          | Female | 15829     | 19183636           | 159789              | 8.2183  | 0.0680             | 43.9476     | 0.2465                 | 15.7914        | 0.1309                 |
|                          | Total  | 33542     | 43651202           | 261536              | 18.7002 | 0.1110             | 100.000     |                        |                |                        |
| Total                    | Male   | 70612     | 111944532          | 162892              | 47.9572 | 0.0417             |             |                        | 100.000        |                        |
|                          | Female | 87664     | 121481295          | 144358              | 52.0428 | 0.0417             |             |                        | 100.000        |                        |
|                          | Total  | 158276    | 233425826          | 239657              | 100.000 |                    |             |                        |                |                        |

Figure 3. Two-way Table of Smoking Status by Sex

| Rao-Scott Chi-Square Test |           |
|---------------------------|-----------|
| Pearson Chi-Square        | 1575.6777 |
| Design Correction         | 1.1923    |
|                           |           |
| Rao-Scott Chi-Square      | 1321.5782 |
| DF                        | 3         |
| Pr > ChiSq                | <.0001    |
|                           |           |
| F Value                   | 440.5261  |
| Num DF                    | 3         |
| Den DF                    | 480       |
| Pr > F                    | <.0001    |
| Sample Size = 158276      |           |

**Figure 4. Chi-Square Goodness-of-Fit Test for No Association**

Figure 1 depicts the “Sum of Weight” (also termed population total). As is depicted, 158,276 adults in the sample represent 233,425,826 adults in the population. The population total is used to compute the percentages displayed in Figure 13. Indeed, suppose we compute the percentage of men among never smokers. If we use the sample counts instead of the population counts, then the proportion will be 40.9%, which is different from the 44.6% provided by SAS. This discrepancy occurs because SAS uses the population counts for computing all percentages. Figure 2 specifies the value of the Fay coefficient used in the analysis: it is 0.5 as is mentioned above. Figure 3 depicts the frequencies and percentages, e.g., among former smokers, 56.1% are male and 43.9% are female. Results depicted in Figure 4 indicate that the association between smoking status and sex is significant, i.e., smoking behaviors among men are different from those among women.

There are additional options in PROC SURVEYFREQ that allow for conducting other types of analyses. One could limit to smokers only when exploring the association between smoking status (occasional smoker, daily smoker) and sex using the “`where SmokStat in (2,3)`” statement. In addition, one could perform a stratified analysis, say by race/ethnicity which is a binary outcome differentiating between non-Hispanic whites and the other racial/ethnic groups, denoted by NHW. First, one needs to sort the data by NHW, then include the “`by NHW`” statement.

## MODEL BUILDING

There is no built-in option for the model selection in the survey package. However, one can use the same logic used in the backward/forward selection methods (Agresti, 2013) and identify the final model based on the SAS output(s). Research goals and study specifics should be used when identifying the initial and final models. For example, in many cases, an investigator already knows what covariates should be controlled for despite their significance or insignificance. In social sciences, sociodemographic characteristics are commonly controlled for and “forced” into a model. The possibility of multi-collinearity should be assessed beforehand. Depending on the study goals, the model building strategies could be rather diverse (Mills, White, Pierce, & Messer, 2011; Soulakova & Crockett, 2016; Soulakova & Danczak, 2017; Soulakova, Li, & Crockett, 2017; Zhang, Martinez-Donate, & Rhoads, 2015).

To illustrate how a logistic regression model can be fitted to the TUS-CPS data we consider estimating the (current) smoking prevalence. Our binary response – current smoking status differentiates between smokers (Yes) and nonsmokers (No) – with the “nonsmoker” being the reference level. Suppose our

study goals are to examine if (1) the relationship between sex (male, female) and smoking prevalence depends on the race/ethnicity (non-Hispanic White, other), and (2) the non-Hispanic whites differ from the other racial/ethnic groups in terms of smoking prevalence. To address these goals, we will fit a model with sex, race/ethnicity, and two-way interaction between the sex, race/ethnicity, and age (as a control variable). Based on the model we will compare the smoking prevalence between non-Hispanic whites and the others.

To fit the model, we use the following SAS code:

```
PROC SURVEYLOGISTIC data=TUS_CPS varmethod=brr(fay);
  class Sex NHW /param=glm;
  model Current(EVENT='Yes') = Age Sex NHW Sex*NHW;
  lsmeans Sex NHW/ilink adj=bon OR CL;
  weight RepWgt0;
  RepWgt RepWgt1-RepWgt160;
RUN;
```

We list all categorical covariates – i.e., sex and race/ethnicity – in the CLASS statement. We use EVENT='Yes' option to request modeling the probability of being a smoker, i.e., treat “No” as a reference level. The LSMEANS statement calls for the overall comparison between non-Hispanic whites and the other racial/ethnic groups while “averaging” over the other effects. The ILINK option requests predicted population margins of the proportions. While we do not perform multiple comparisons, we use the ADJ=BON option to request the 95% confidence interval for the difference between the proportion of non-Hispanic white smokers and the proportion of the other smokers. The OR and CI options request the odds ratio(s) and the confidence interval(s), respectively. Figure 5 – Figure 8 display the key parts of the output.

| Testing Global Null Hypothesis: BETA=0 |            |    |            |
|----------------------------------------|------------|----|------------|
| Test                                   | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio                       | 2532940.49 | 4  | <.0001     |
| Score                                  | 2408537.71 | 4  | <.0001     |
| Wald                                   | 1016.7036  | 4  | <.0001     |

Figure 5. Model Fit Information

| Type 3 Analysis of Effects |    |                 |            |
|----------------------------|----|-----------------|------------|
| Effect                     | DF | Wald Chi-Square | Pr > ChiSq |
| Age                        | 1  | 441.7965        | <.0001     |
| Sex                        | 1  | 321.641         | <.0001     |
| NHW                        | 1  | 308.5348        | <.0001     |
| Sex*NHW                    | 1  | 116.5608        | <.0001     |

Figure 6. Significance of Covariates

| NHW Least Squares Means |          |                |         |         |       |         |         |         |                        |            |            |
|-------------------------|----------|----------------|---------|---------|-------|---------|---------|---------|------------------------|------------|------------|
| NHW                     | Estimate | Standard Error | z Value | Pr >  z | Alpha | Lower   | Upper   | Mean    | Standard Error of Mean | Lower Mean | Upper Mean |
| Non-Hispanic White      | -1.7801  | 0.01036        | -171.74 | <.0001  | 0.05  | -1.8004 | -1.7598 | 0.1443  | 0.001280               | 0.1418     | 0.1468     |
| Other                   | -2.2117  | 0.02051        | -107.84 | <.0001  | 0.05  | -2.2519 | -2.1715 | 0.09871 | 0.001824               | 0.09519    | 0.1023     |

**Figure 7. The Model-based Smoking Prevalence Estimates for Non-Hispanic White and the Other Racial/ethnic Groups**

| Differences of NHW Least Squares Means<br>Adjustment for Multiple Comparisons: Bonferroni |       |          |                |         |         |        |       |        |        |           |           |            |                                       |                                       |                      |                      |
|-------------------------------------------------------------------------------------------|-------|----------|----------------|---------|---------|--------|-------|--------|--------|-----------|-----------|------------|---------------------------------------|---------------------------------------|----------------------|----------------------|
| NHW                                                                                       | NHW   | Estimate | Standard Error | z Value | Pr >  z | Adj P  | Alpha | Lower  | Upper  | Adj Lower | Adj Upper | Odds Ratio | Lower Confidence Limit for Odds Ratio | Upper Confidence Limit for Odds Ratio | Adj Lower Odds Ratio | Adj Upper Odds Ratio |
| Non-Hispanic White                                                                        | Other | 0.4316   | 0.02457        | 17.57   | <.0001  | <.0001 | 0.05  | 0.3834 | 0.4798 | 0.3834    | 0.4798    | 1.540      | 1.467                                 | 1.616                                 | 1.467                | 1.616                |

**Figure 8. Comparison of Odds of Being a Smoker for non-Hispanic Whites Relative to the Other Racial/ethnic Groups**

**Error! Reference source not found.**Figure 5 illustrates that the model is significant. Figure 6 depicts significance of the interaction and the other factors. Thus, we conclude that the smoking behaviors jointly “depend” on sex and race/ethnicity, i.e., among non-Hispanic whites, and that the effect of sex on smoking differs from the one among the other racial/ethnic groups. Figure 7 depicts the smoking prevalence estimates for non-Hispanic whites (14.43%, SE=0.13%, CI=14.18%:14.68%) and the others (9.87%, SE=0.18%, CI=9.52%:10.23%). Figure 8 illustrates that non-Hispanic whites have higher odds of being a smoker relative to the “other” racial/ethnic group (Odds Ratio=1.54,  $p<0.0001$ , 95% CI=1.47:1.62).

## CONCLUSION

The paper discusses two SAS®9.4 procedures for analyses of complex survey data. Applications to the 2014-15 TUS-CPS are discussed. The TUS-CPS provides a rich source of data on smoking behaviors for diverse populations in the US. After completing the data preparation task, the data can be relatively easy analyzed using SAS® survey package. The package contains many other analytical approaches, e.g., the surveymeans and surveyreg procedures, which can be used similarly to the methods discussed in this paper.

## REFERENCES

- Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley-Interscience.
- Fay, R. E., & Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In *Proceedings of the Section on Government Statistics, American Statistical Association, Alexandria, VA* (pp. 154–159).
- Judkins, D. (1990). Fay’s Method for Variance Estimation. *Journal of Official Statistics*, 6(3), 223–239.
- Mills, A. L., White, M. M., Pierce, J. P., & Messer, K. (2011). Home smoking bans among U.S. households with children and smokers: Opportunities for intervention. *American Journal of Preventive Medicine*, 41(6), 559–565.
- Rao, J. N. K., & Scott, A. J. (1981). The Analysis of Categorical Data From Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association*, 76(374), 221.

- Soulakova, J. N., & Crockett, L. J. (2016). Unassisted Quitting and Smoking Cessation Methods Used in the United States: Analyses of 2010–2011 Tobacco Use Supplement to the Current Population Survey Data. *Nicotine & Tobacco Research*, 242(5), ntw273.
- Soulakova, J. N., & Danczak, R. R. (2017). Impact of Menthol Smoking on Nicotine Dependence for Diverse Racial/Ethnic Groups of Daily Smokers. *Healthcare (Basel, Switzerland)*, 5(1).
- Soulakova, J. N., Li, J., & Crockett, L. J. (2017). Race/ethnicity and intention to quit cigarette smoking. *Preventive Medicine Reports*, 5, 160–165.
- U.S. Bureau of Labor Statistics, & U.S. Census Bureau. (2006). Design and Methodology: Current Population survey, Technical Paper 66. Retrieved from <https://www.census.gov/programs-surveys/cps/technical-documentation/complete.html>
- US Department of Commerce, & Census Bureau. (2016). National Cancer Institute and Food and Drug Administration co-sponsored Tobacco Use Supplement to the Current Population Survey. 2014-15. Retrieved from [http://thedataweb.rm.census.gov/ftp/cps\\_ftp.html#cpssupps](http://thedataweb.rm.census.gov/ftp/cps_ftp.html#cpssupps)
- Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). Springer.
- Zhang, X., Martinez-Donate, A., & Rhoads, N. (2015). Parental Practices and Attitudes Related to Smoke-Free Rules in Homes, Cars, and Outdoor Playgrounds in US Households With Underage Children and Smokers, 2010-2011. *Preventing Chronic Disease*, 12, E96.

## ACKNOWLEDGMENTS

We are thankful to James Holland for proofreading the paper.

## CONTACT INFORMATION

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

Dr. Trung Ha,  
Burnett School of Biomedical Sciences, College of Medicine,  
University of Central Florida,  
6900 Lake Nona Blvd, Orlando, FL 32827  
[trung.ha@ucf.edu](mailto:trung.ha@ucf.edu)

Dr. Julia N. Soulakova,  
Burnett School of Biomedical Sciences, College of Medicine,  
University of Central Florida,  
6900 Lake Nona Blvd, Orlando, FL 32827  
[julia.soulakova@ucf.edu](mailto:julia.soulakova@ucf.edu)

Website: <https://med.ucf.edu/biostatistics/>