2018-2019 Tobacco Use Supplement to the Current Population Survey (TUS-CPS) User Guide for Conducting Weighted Analyses

Created February 2021

Table of Contents

Introduction

Purpose

This document includes specific information and guidance on conducting weighted analyses using 2018-2019 TUS data. Developed as a supplement to the 2018-2019 technical documentation, this guide provides illustrative examples of different types of analyses, including statistical code in both SAS and SAS-Callable SUDAAN. A working knowledge of either SAS or SAS-Callable SUDAAN will be helpful in following these examples, which are intended primarily to demonstrate the necessary coding specifications (e.g., replicate weighting) for obtaining properly weighted statistics and appropriate variance estimates using the 2018-2019 dataset. When working with earlier waves of TUS-CPS data, users should consult the technical documentation specific to those waves. The scenarios included here are by no means exhaustive; rather, they are intended to serve as an orientation to working with TUS-CPS weighted data. For greater detail on survey design and methodology, please refer to the 2018-2019 technical documentation, particularly the Source and Accuracy Statement (Attachment 16), and other resources on the CPS and TUS-CPS websites. More information can be found below.

What is the Current Population Survey?

The Current Population Survey (CPS) is valuable source of information on work, earnings, occupation, and education in the United States. Continuously administered by the U.S. Census Bureau on behalf of the U.S. Bureau of Labor Statistics, CPS is the primary source of data on unemployment and the labor force. Approximately 54,000 households are interviewed for CPS each month. Selected households complete the survey once a month for four consecutive months in one year, then are interviewed again for four corresponding months a year later. This ensures that the sample is continually replenished. Each month, a new group of households begins the rotation (four months on/eight months off/four months on), while another group of households concludes its rotation. This design also ensures that there is substantial continuity from month to month and year to year.

Households are selected based on a two-stage sampling scheme: sample housing units are chosen from selected primary sampling units (PSUs). The CPS sampling frame generally comes from the Master Address File, which is managed by the U.S. Census Bureau and updated by the U.S. Postal Service, among others. Within the household, U.S. residents aged 15 or over who are not in the armed forces and are not institutionalized are eligible to participate in the CPS Core survey. Generally, one person responds for all eligible household members, which means that some proxy data may be collected.

For a more in-depth description of the sampling methodology and study design, please see the following resources: 2018-2019 TUS-CPS technical documentation and CPS Technical Paper 77.

What is the Tobacco Use Supplement to the Current Population Survey?

The Tobacco Use Supplement to the Current Population Survey (TUS-CPS) is a unique and valuable data source that can be used to track long-term tobacco use trends, evaluate programs, and analyze tobacco-related research questions (e.g., tobacco-related health disparities and effects of tobacco control policies) at the

national, state, and county[1] levels. Since it was first launched in 1992, the TUS-CPS has been conducted every three to four years. Within most waves, the TUS-CPS is administered during three separate "survey months," typically spaced four months apart in order to minimize the number of individuals selected and surveyed more than once within the full-wave sample. The 2018-2019 TUS-CPS wave includes data collected in July 2018, January 2019, and May 2019. NCI has sponsored the TUS since its inception, with co-sponsorship from the U.S. Food and Drug Administration (FDA) since 2014 and the Centers for Disease Control and Prevention (CDC) from 2001-2007. For more information, visit cancercontrol.cancer.gov/tus-cps.

As a supplement to the CPS, TUS-CPS uses the sampling methodology of its parent survey, with some additional criteria. Currently, individuals aged 18 years and older who have completed the CPS Core survey are eligible for participation in TUS-CPS. In households with only one or two TUS-eligible members, all eligible individuals are selected for self-interview. In households with three or four TUS-eligible members, however, two of those individuals are randomly selected for self-interview; in households with more than four TUS-eligible members, three of those individuals are randomly selected. Proxy responses are collected only when an individual is selected for self-interview and cannot be reached after four attempts. In waves prior to 2018-2019, TUS-CPS methodology differed slightly in certain respects. For example, early waves included interviews with 15- to 17-year-olds, and, beginning with the 2014-2015 wave, the number of self-response interviews per household was reduced to decrease household response burden. Please refer to the technical documentation for specifics if working with any earlier waves.

To obtain TUS-CPS data files, SAS programs to import files, and technical documentation for all waves, including 2018-2019, please visit the TUS-CPS Questionnaires and Data Files webpage. For questions about the TUS-CPS, contact us at ncidccpsbrpadvances@mail.nih.gov.

Analyzing TUS-CPS Weighted Data

TUS-CPS includes full-sample weights and replicate weights[2]. Full sample weights are created to compensate for differential selection probabilities, nonresponse, and under-coverage of the target population of U.S. adults. Replicate weights, which can accommodate various types of statistical analyses, are created to more accurately estimate standard errors by accounting for the complex survey design. More information about the construction of these weights can be found in the CPS Technical Paper 77 and the TUS-CPS Technical Documentation.

In short, the process of creating the TUS full sample weights begins with the final CPS weights, which control for the probability of selection into CPS and a response to CPS. To create the TUS nonresponse weights, the CPS weights are adjusted to account for supplement nonresponses (including both self- and proxy-nonresponses). Separately, to create the TUS self-nonresponse weights, the CPS weights are adjusted to account for supplement self-nonresponses only. Next, the two sets of nonresponse-adjusted weights are post-stratified to match the state-specific Black alone/non-Black alone population distribution from the 2010 Census, then to match the national demographic totals for age by sex by race by origin groups, and, finally, to match the state-

[1] County-level data not available for all counties.
[2] Replicate weights are not included in the data files on the TUS-CPS website. To obtain the latest replicate weights, please visit https://www.census.gov/data/datasets/time-series/demo/cps/cps-supp_cps-repwgt/cps-tobacco.2018.html. Replicate weight files prior to 2010 are available upon request from NCI.

specific population totals for the age by sex by race groups, which are derived from the Census Bureau's Population Estimates Program.

TUS-CPS replicate weights are generated by the U.S. Census Bureau using algorithms that simulate drawing a set of replicated subsamples with the same sampling design as the full sample. Then, by measuring the variation between estimates produced using the subsamples and the full sample, statistical programs can compute design-adjusted standard errors to accompany weighted point estimates. There are several different variance estimation methods that use this replicated subsampling approach to create and use replicate weights. These approaches differ in how the subsamples are generated and in the formulae used to measure variation across the replicated estimates. Methods include, among others, balanced repeated replication (BRR), jackknife, and bootstrap (Wolter, 2007). A variant of BRR, Fay's method, is used in variance estimation for TUS-CPS and is demonstrated in the SAS examples that follow. In addition, see Box A for an illustration of the use of replicate weights and Fay's method to estimate standard error.

Box A. Using replicate weights and Fay's method to estimate standard error

Let $\hat{Y}$ be the weighted estimate (based on the full sample weight) for a given statistic Y (e.g., total, mean, regression coefficient) for an outcome of interest (e.g., point prevalence of current smoking). The standard error associated with $\hat{Y}$ based on the Fay-adjusted BRR method can be calculated as:

$$SE(\hat{Y}) = \sqrt{\frac{4}{R}\sum_{r=1}^{R}\left(\hat{Y}_{(r)} - \hat{Y}\right)^2},$$

where R is the total number of replicate weights, $\hat{Y}_{(r)}$ is the estimate of Y based on the r-th replicate weight, $r = 1, \ldots, R$. The constant $\frac{4}{R}$ is specific to the Fay-adjusted BRR method. A different constant would be used if weights were based on a different replication method. The total number of replicates R varies for TUS-CPS depending on the survey wave. For the 2018-2019 wave, R=160.

Next, we provide a simple illustration on how to use the standard error formula. Assume that an example survey has $R = 3$, $\hat{Y} = 10, \hat{Y}_{(1)} = 8, \hat{Y}_{(2)} = 11$, and $\hat{Y}_{(3)} = 12$. Then:

$$SE(\hat{Y}) = \sqrt{\frac{4}{3}((8-10)^2 + (11-10)^2 + (12-10)^2)} = 3.46.$$

A more detailed explanation of these formulae can be found in a 2013 webinar, Statistical Aspects on the Best Practices for Analyzing TUS-CPS Data.

The remainder of this document will highlight several different examples, provide corresponding code and output, and describe related considerations.

Basic Weighted Analyses

Obtaining Frequencies and Percentages

Example 1: Calculating current smoking prevalence among self-respondents for the 2018-2019 wave (July 2018, January 2019, and May 2019)

For the 2018-2019 wave of TUS-CPS, there are several available files to incorporate: one survey data file for each month (July 2018, January 2019, and May 2019); a SAS program to read in the survey data files; two replicate weight files for each month (one for nonresponse replicate weights and one for self-response replicate weights); and two SAS programs to read in the replicate weights (one for nonresponse replicate weights and one for self-response replicate weights). There are 160 replicate weights in each of the replicate weight files. The unique identifiers for 2018-2019 include a unique household identifier (QSTNUM) and a unique person identifier (OCCURNUM).

The following steps outline data preparation for all analyses that use the full 2018-2019 wave:

1. Create SAS libraries that are associated with the folders containing survey data and replicate weight files. Note that the libraries and file paths below are examples and should be changed to match the location where the files are saved.

```
Libname Survey   "T:\Temp\CPS\survey\";
Libname RepFiles "T:\Temp\CPS\repwgts\";
```

2. Read in the main survey data files for each month (July 2018, January 2019, and May 2019), separately, using the SAS programs provided on the TUS-CPS website: https://cancercontrol.cancer.gov/brp/tcrb/tus-cps/questionnaires.html. Code for this step is not included in the Appendix.

3. For each month, select TUS-CPS self-respondents. For example, for July 2018:

```
/* Read the main survey files for each month (July 2018, January 2019, May
2019)  */
Data CPSJul18;
  Set Survey.CPSJul18;
  If PRPerTyp=2;  /* Adult Civilian Household Member (15+ years old) */
  If Intrview=1;  /* TUS-CPS Interviews */
  If PRS64=1;     /* Self-Respondents */
Run;
```

4. For each month, sort the survey data files by the unique identifiers (QSTNUM, OCCURNUM). For example, for July 2018:

```
/* Sort the main survey data for each month */
Proc Sort Data=CPSJul18;
  By QstNum OccurNum;
Run;
```

5. Read in the self-response replicate weight files for each month using the SAS code, called "input statements," provided on the Census website. For January and May 2019: https://www.census.gov/data/datasets/time-series/demo/cps/cps-supp_cps-repwgt/cps-tobacco.html; for July 2018: https://www.census.gov/data/datasets/time-series/demo/cps/cps-supp_cps-repwgt/cps-tobacco.2018.html. Code for this step is not included in the Appendix.

6. For each month, sort the replicate weight data files by the unique identifiers (QSTNUM, OCCURNUM). For example, for July 2018:

```
/* Sort self-response replicate weights */
Proc Sort Data=repjul18;
  By QstNum OccurNum;
Run;
```

7. For each month, merge the survey data and replicate weight files by the unique identifiers. For example, for July 2018:

```
/* Merge survey and replicate weight data */
Data cpsjul18;
  Merge cpsjul18(In=In1)
        repjul18(In=In2);
  By QstNum OccurNum;
  If In1;
  If Not In2 Then Put "Error in Merge: " QstNum= OccurNum=; /*Optional check
for records in the main survey data file that are not in the replicate weight
file; outputs message to SAS log*/
Run;
```

8. Concatenate (stack) the three datasets.

```
/* Concatenate the three datasets */
Data CPS1819;
  Set cpsjul18
      cpsjan19
      cpsmay19;
Run;
```

9. Divide person-weights and replicate weights by three (the number of months of data being combined for analysis). Otherwise, the dataset will yield a sample three times the size of the U.S. population.

```
/* Divide weights by number of months (surveys) being combined */
/* PWSRWgt is the weight from the main survey file.             */
/* RepWt0 is the same as PWSRWgt                                */
/* RepWgt1-RepWgt160 are the replicate weights                  */
Data CPS1819;
  Set CPS1819;
  Array Wgts(162) PWSRWgt RepWgt0-RepWgt160;
  Do I = 1 to 162;
    Wgts(I)=Wgts(I)/3;
      End;
Run;
```

10. Recode variables for analysis, in this case current cigarette smoking status (yes/no, called CURRSMK); cigarettes smoked per day (continuous, among current smokers, called CIGPD); sex (male/female, called SEX); age group (18-24/25-34/35-44/45-54/55-64/65+, called AGEGRP); and race/ethnicity (non-Hispanic White/non-Hispanic Black/Hispanic/Other, including multiracial, called RACEETH). The record layout ("codebook") can be found in Attachment 7 of the Technical Documentation corresponding to each month.

Tobacco Use Supplement
Current Population Survey

```
Data CPS1819;
  Set CPS1819;
  /* CurrSmk: Current Cigarette Smoking Status */
  If PEA1=1 & PEA3 In (1,2) Then CurrSmk=1; /* Current Cigarette Smoker */
  Else If PEA1=1 & PEA3=3 Then CurrSmk=0;   /* Former Cigarette Smoker */
  Else If PEA1=2 Then CurrSmk=0;            /* Never Cigarette Smoker */
  Else CurrSmk=.;

  /*CurrVape: Current E-Cigarette Use*/
  If PEJ1A3_5=1 & PEJ2A3_5 in (1,2) then CurrVape=1;  /*Current*/
  Else If PEJ1A3_5=1 & PEJ2A3_5=3 then CurrVape=0;    /*Ever used, not
  currently vaping*/
  Else if PEJ1A3_5=2 Then CurrVape=0; /*Never vaped*/
  Else CurrVape=.;

  /* CigPD: Number of Cigarettes Per Day */
   If PEA1=1 & PEA3=1 & (1<=PTB1<=99) Then CigPD=PTB1;      /* Daily Smokers
  */
   Else If PEA1=1 & PEA3=2 & (1<=PEC1<=30) & (1<=PTC1A<=99) Then
   CigPD=(PEC1/30)*PTC1A; /* Non-Daily Smokers */
  Else CigPD=.;

  /* Sex: Sex */
  If PESex=1 Then Sex=1;        /* Male */
  Else If PESex=2 Then Sex=0;   /* Female */

  /* AgeGrp: Age Group */
  If (18<=PrtAge<=24) Then AgeGrp=1;       /* 18-24 */
  Else If (25<=PrtAge<=34) Then AgeGrp=2;  /* 25-34 */
  Else If (35<=PrtAge<=44) Then AgeGrp=3;  /* 35-44 */
  Else If (45<=PrtAge<=54) Then AgeGrp=4;  /* 45-54 */
  Else If (55<=PrtAge<=64) Then AgeGrp=5;  /* 55-64 */
  Else If (PrtAge=>65) Then AgeGrp=6;      /* 65+ */
  Else AgeGrp=.;

  /* RaceEthn: Race/Ethnicity */
  If PEHSPNON=1 Then RaceEthn=3;    /* Hispanic */
  Else If PEHSPNON=2 & PTDTRace=1 Then RaceEthn=1;  /* Non-Hispanic, White
  Alone */
  Else If PEHSPNON=2 & PTDTRace=2 Then RaceEthn=2;  /* Non-Hispanic, Black
  Alone */
  Else RaceEthn=4;  /* Non-Hispanic, All other (including multiracial) */

  Label CurrSmk  = "Current Cigarette Smoking Status"
            CurrVape = "Current E-Cigarette Use"
            CigPD    = "Number of Cigarettes Per Day"
            AgeGrp   = "Ages"
            RaceEthn = "Race/Ethnicity";
  Keep Intrview PRS64 CurrSmk CurrVape CigPD Sex AgeGrp RaceEthn PWSRWgt
  RepWgt1-RepWgt160;
  Format CurrSmk CurrSmkF. CurrVape CurrVapeF. Sex SexF. AgeGrp AgeGrpF.
  RaceEthn RaceEthF.;
Run;
```

The following code will output frequency tables for current cigarette smoking status by sex and age group among all adult self-respondents (that is, excluding proxy responses) using SAS and SAS-Callable SUDAAN. Both statistical programs will produce the same estimates, which will be weighted using the self-response person-weight (PWSRWGT) and self-response replicate weights (REPWGT1-REPWGT160). It is not necessary to run both programs.

SAS:

```
/* SAS Code for Generating Estimates for Table 1 */
Proc SurveyFreq Data=CPS1819 VarMethod=BRR (Fay=0.5);
  Tables CurrSmk / cl;
  Tables (Sex AgeGrp)*CurrSmk / cl row;
  Weight PWSRWgt;
  RepWeights RepWgt1-RepWgt160;
Run;
```

SUDAAN:

```
Proc Crosstab Data=CPS1819 Design=BRR;
  Weight PWSRWgt;
  RepWgt RepWgt1-RepWgt160/ADJFay=4;
  Class CurrSmk Sex AgeGrp;
  Tables (Sex AgeGrp)*CurrSmk;
  Print NSum="Sample Size"
        WSum="Weighted Size"
        RowPer="Percent"
        SERow="SE Error"
        LowRow="Lower 95% CI"
        UpRow="Upper 95% CI"
        /Style=NCHS NSumFmt=F8.0 WSumFmt=F10.0 RowPerFmt=F6.1 SERowFmt=F5.1
        LowRowFmt=F6.1 UpRowFmt=F6.1;
Run;
```

The output generated using this code can be reorganized into the following table:

Table 1: Current Cigarette Smoking Status by Sex and Age Group Among Adult (18+ years) Self-Respondents

	Overall		Current Cigarette Smoking: Yes				Current Cigarette Smoking: No			
	Sample	Population	Percent	Sample	Population	95% CI[1]	Percent	Sample	Population	95% CI[1]
Overall	136806	248831743	11.4	16570	28391408	11.2-11.6	88.6	120236	220440335	88.4-88.8
Sex										
Male	62162	119859783	12.9	8368	15510251	12.6-13.3	87.1	53794	104349532	86.7-87.4
Female	74644	128971960	10.0	8202	12881157	9.7-10.3	90.0	66442	116090803	89.7-90.3
Age										
18-24	7616	29477095	7.4	719	2186777	6.7-8.2	92.6	6897	27290318	91.8-93.3
25-34	21547	44562597	12.2	2738	5417315	11.7-12.7	87.8	18809	39145282	87.3-88.3
35-44	22340	40566024	12.9	3098	5240917	12.4-13.5	87.1	19242	35325107	86.5-87.6
45-54	20861	40661567	13.6	3102	5543159	13.1-14.2	86.4	17759	35118409	85.8-86.9
55-64	25896	41929841	14.9	4058	6265996	14.4-15.5	85.1	21838	35663844	84.5-85.6
65+	38546	51634620	7.2	2855	3737244	6.9-7.6	92.8	35691	47897376	92.4-93.1

Source: TUS-CPS, 2018-2019

[1] Wald confidence limits

Obtaining Means and Standard Errors
Example 2: Calculating mean cigarettes per day among adult smokers during the 2018-2019 wave (self-response)

Some research questions may require that analyses are conducted within a particular subset of respondents. For example, if a researcher plans to calculate the mean number of cigarettes smoked per day among current smokers, respondents who are former or never smokers would not be of interest because they most likely do not have a valid response to the cigarettes per day item. When performing subgroup analyses on data collected using simple random sampling, it may be acceptable to drop cases from subgroups that are not of interest. However, when working with complex survey data such as TUS-CPS, it is important to keep all respondents in the dataset throughout analysis (whether of frequency-based or continuous measures). Subsetting may result in a loss of information on the complex study design and a miscalculation of the variance because statistical computing programs will not have, or will not take into consideration, the sampling information for all observations. Instead, proper methods for conducting subgroup analyses incorporate sampling information for all observations, even those that are not in the subpopulation of interest. For more information, please see Paper 449-2013 from the SAS Global Forum 2013.

To conduct subgroup analyses using SAS SURVEY procedures, researchers should use DOMAIN statements.[3] To conduct subgroup analyses using SUDAAN, researchers should use SUBPOPULATION statements. In both SAS and SUDAAN, use of WHERE statements to subset the data within a procedure should be avoided. Additionally, users should not subset data in a separate procedure prior to running analyses, as this can also lead to incorrect results and errors in the procedure.

To illustrate this, the following example will calculate the mean number of cigarettes smoked per day among all adult smokers and among adult smokers by sex. The following code will use the dataset prepared in Example 1 to output means, standard errors, and 95% confidence intervals for cigarettes smoked per day (CIGPD) by sex (SEX) among current cigarette smoking (CURRSMK=1) adult self-respondents. Estimates will be weighted using the adjusted self-response person-weight (PWSRWGT) and self-response replicate weights (REPWGT1-REPWGT160).

SAS:

```
/* SAS Code for Generating Estimates for Table 2 */
Proc SurveyMeans Data=CPS1819 VarMethod=BRR (Fay=0.5);
  Var CigPD;
  Domain CurrSmk;
  Weight PWSRWgt;
  RepWeights RepWgt1-RepWgt160;
Run;

Proc SurveyMeans Data=CPS1819 VarMethod=BRR (Fay=0.5);
  Var CigPD;
  Domain CurrSmk*Sex;
  Weight PWSRWgt;
  RepWeights RepWgt1-RepWgt160;
Run;
```

[3] Some SURVEY procedures such as PROC SURVEYFREQ do not use a DOMAIN statement. In PROC SURVEYFREQ, the combination of variables specified and joined by asterisks in the TABLES statement can specify the subset of interest while preserving survey weights. Please see SAS guidance for more details.

SUDAAN:

```
Proc Descript Data=CPS1819 Design=BRR;
  Weight PWSRWgt;
  RepWgt RepWgt1-RepWgt160/ADJFay=4;
  Class Sex/NoFreq;
  Var CigPD;
  SubPopN CurrSmk=1/Name="Current Cigarette Smokers";
  Print NSum="Sample Size"
       WSum="Weighted Size"
       Mean="Mean"
       SEMean="Std Error"
       LowMean="Lower 95% CI"
       UpMean="Upper 95% CI"
       /Style=NCHS NSumFmt=F8.0 WSumFmt=F10.0 MeanFmt=F6.1 SEMeanFmt=F5.1
       LowMeanFmt=F6.1 UpMeanFmt=F6.1;
 Run;
```

The output generated using this code can be reorganized into the following table:

Table 2: Cigarettes Smoked Per Day by Sex Among Currently Smoking Adult (18+ years) Self-Respondents

	Cigarettes Per Day		
	Mean	Standard Error	95% CL for Mean
Overall	11.4	0.1	11.3-11.6
Sex			
Male	11.9	0.1	11.7-12.2
Female	10.8	0.1	10.6-11.1

Source: TUS-CPS, 2018-2019

Regression Analyses

Example 3. Using binomial logistic regression to estimate the adjusted odds ratio for the association between smoking status and current e-cigarette use

To prepare the data for analysis, follow the preliminary steps laid out in Example 1. The following code will output logistic regression coefficients, design-adjusted standard errors and confidence intervals, and corresponding odds ratios for the association between current e-cigarette use (yes/no, CURRVAPE) and each model parameter: sex, age group, race/ethnicity, and current smoking status. Estimates will be weighted using the adjusted self-response person-weight (PWSRWGT) and self-response replicate weights (REPWGT1-REPWGT160).

SAS:

```
/* SAS Code for Generating Estimates for Table 2 */
Proc SurveyLogistic Data=CPS1819 VarMethod=brr (Fay=0.5);
  Class Sex (ref='Female') AgeGrp (ref='18-24') RaceEthn (ref='Non-Hispanic,
  White Alone') CurrSmk (ref='Not Current') / Param=Ref;
  Model CurrVape = Sex AgeGrp RaceEthn Currsmk;
  Weight PWSRWgt;
  Repweights RepWgt1-RepWgt160;
Run;
```

SUDAAN:

```
Proc RLogist Data=CPS1819 Design=BRR;
  Weight PWSRWGT;
  RepWgt REPWGT1-REPWGT160/ADJFay=4;
  Class Sex AgeGrp RaceEthn CurrSmk;
  Reflevel Sex=0 AgeGrp=1 RaceEthn=1 CurrSmk=0;
  Model CurrVape = Sex AgeGrp RaceEthn CurrSmk ;
Run;
```

The output generated using this code can be reorganized into the following table:

Table 3: Adjusted Odds Ratios for the Association between Model Parameters and Current Vaping Among Adult (18+ years) Self-Respondents

		Adjusted Odds Ratio	95% Confidence Interval
Sex			
	Male	1.43	1.31-1.56
	Female	Ref.	Ref.
Age			
	18-24	Ref.	Ref.
	25-34	0.55	0.47-0.64
	35-44	0.33	0.28-0.40
	45-54	0.24	0.20-0.28
	55-64	0.19	0.16-0.22
	65+	0.09	0.08-0.12
Race/Ethnicity			
	Non-Hispanic White	Ref.	Ref.
	Hispanic	0.46	0.40-0.55
	Non-Hispanic Black	0.33	0.27-0.41
	Other, including multiracial	0.63	0.51-0.77
Cigarette Use			
	Current Smoker	5.51	4.99-6.09
	Non-Smoker	Ref.	Ref.

Source: TUS-CPS, 2018-2019

State-Level Analyses

One of the most valuable aspects of TUS-CPS is its ability to support estimates at the state level and for some metropolitan areas. For some examples of research that uses TUS-CPS state-level data, please see the following references:

- Jemal, A., Thun, M., Yu, X.Q. et al. (2011). Changes in smoking prevalence among U.S. adults by state and region: Estimates from the Tobacco Use Supplement to the Current Population Survey, 1992-2007. *BMC Public Health*, vol. 11, no. 512.
- Farrelly, M.C., Arnold, K.Y., Juster, H.R., and Allen, J.A. (2014). Quantifying the Effect of Changes in State-Level Adult Smoking Rates on Youth Smoking. *Journal of Public Health Management and Practice*, vol. 20, no. 2.
- Dahne, J., Wahlquist, A.E., Garrett-Mayer, E., Heckman, B.W., Cummings, K.M., and Carpenter, M.J. (2018). State Tobacco Policies as Predictors of Evidence-Based Cessation Method Usage: Results from a Large, Nationally Representative Dataset. *Nicotine & Tobacco Research*, vol. 20, no. 11.

Although each month of TUS-CPS is weighted to the U.S. population, when conducting descriptive analyses of the data at the state level or smaller, it is typically necessary to use the full wave of data (all three months) and adjust the weights accordingly (as demonstrated above). This ensures a sufficient sample size at the state level. In some cases, it may be acceptable to analyze data within a single, larger state, for example, when comparing estimates between survey months before or after the implementation of a new policy. For some more powerful analyses, such as regression analyses, use of only one survey month may be adequate.

Conclusion

This document provides guidance on, and examples of, weighted analyses using 2018-2019 TUS data. It is primarily intended to demonstrate the necessary coding specifications (e.g., replicate weighting) for obtaining properly weighted statistics and appropriate variance estimates using the 2018-2019 dataset. When working with earlier waves of TUS-CPS data, users should consult the technical documentation specific to those waves.

The scenarios included are by no means exhaustive; rather, they are intended to serve as an orientation to working with TUS-CPS weighted data. For greater detail on survey design and methodology, please refer to the TUS-CPS technical documentation, particularly the Source and Accuracy Statement (Attachment 16), available at cancercontrol.cancer.gov/brp/tcrb/tus-cps/questionnaires-data, as well as other resources on the CPS and TUS-CPS websites (see Resources and Links section below). Additional questions and/or suggested additions to future guidance documents may be directed to ncidccpsbrpadvances@mail.nih.gov.

Resources and Links

Additional information, SAS code, and examples of TUS data analysis can be found on the TUS-CPS website:

- TUS-CPS Frequently Asked Questions Page
- 2020 Informational Session Questions and Responses
- 2009 Tobacco Use Supplement to the Current Population Survey (TUS-CPS) Users' Workshop – Presentation on *Tips and Tricks of Handling the TUS Data*

- 2013 TUS-CPS Webinar – Presentation on *Tips and Tricks of Handling TUS-CPS Data*
- 2007 Overlap Sample Report

Other helpful references for analysis of survey data may include:

- Dahne, J., Wahlquist, A.E., Garrett-Mayer, E., Heckman, B.W., Cummings, K.M., and Carpenter, M.J. (2018). State Tobacco Policies as Predictors of Evidence-Based Cessation Method Usage: Results from a Large, Nationally Representative Dataset. *Nicotine & Tobacco Research*, vol. 20, no. 11.
- Farrelly, M.C., Arnold, K.Y., Juster, H.R., and Allen, J.A. (2014). Quantifying the Effect of Changes in State-Level Adult Smoking Rates on Youth Smoking. *Journal of Public Health Management and Practice*, vol. 20, no. 2.
- Heeringa, S.G., West, B.T., and Berglund, P.A. (2017). *Applied Survey Data Analysis* (2nd ed.). Boca Raton, Florida: CRC Press.
- Jemal, A., Thun, M., Yu, X.Q. et al. (2011). Changes in smoking prevalence among U.S. adults by state and region: Estimates from the Tobacco Use Supplement to the Current Population Survey, 1992-2007. *BMC Public Health*, vol. 11, no. 512.
- Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*. USA: John Wiley & Sons.
- Lee, E.S. and Forthofer, R.N. (2005). Analyzing Complex Survey Data (2nd ed.). *Quantitative Applications in the Social Sciences*, vol. 71. Thousand Oaks, California: Sage Publications.

Additional questions may be directed to ncidccpsbrpadvances@mail.nih.gov.

References

Wolter, K. (2007). *Introduction to Variance Estimation* (2nd ed.). New York: Springer-Verlag.


```
Libname Survey   "T:\Temp\CPS\survey\";
Libname RepFiles "T:\Temp\CPS\repwgts\";

Proc Format;
  Value CurrSmkF
    0 = "Not Current"
      1 = "Current"
      ;
  Value CurrVapeF
    0 = "Not Current"
      1 = "Current"
      ;
  Value SexF
    0 = "Female"
      1 = "Male"
      ;
  Value AgeGrpF
    1 = "18-24"
      2 = "25-34"
      3 = "35-44"
      4 = "45-54"
      5 = "55-64"
      6 = "65+"
      ;
  Value RaceEthF
    1 = "Non-Hispanic, White Alone"
      2 = "Non-Hispanic, Black Alone"
      3 = "Hispanic"
      4 = "Non-Hispanic, Other"
      ;
run;

/* Read the main survey files for each month (July 2018, January 2019, May 2019)  */
Data CPSJul18;
  Set Survey.CPSJul18;
  If PRPerTyp=2;  /* Adult Civilian Household Member (15+ years old) */
  If Intrview=1;  /* TUS-CPS Interviews */
  If PRS64=1;     /* Self-Respondents */
Run;

Data CPSJan19;
  Set Survey.CPSJan19;
  If PRPerTyp=2;  /* Adult Civilian Household Member (15+ years old) */
  If Intrview=1;  /* TUS-CPS Interviews */
  If PRS64=1;     /* Self-Respondents */
Run;

Data CPSMay19;
  Set Survey.CPSMay19;
  If PRPerTyp=2;  /* Adult Civilian Household Member (15+ years old) */
  If Intrview=1;  /* TUS-CPS Interviews */
  If PRS64=1;     /* Self-Respondents */
Run;
```


```
/* Sort the main survey data for each month */
Proc Sort Data=CPSJul18;
  By QstNum OccurNum;
Run;

Proc Sort Data=CPSJan19;
  By QstNum OccurNum;
Run;

Proc Sort Data=CPSMay19;
  By QstNum OccurNum;
Run;

/*  Read self response replicate weights */
Data repjul18;
  Set RepFiles.RepWgts_Jul18;
Run;

Data repjan19;
  Set RepFiles.RepWgts_Jan19;
Run;

Data repmay19;
  Set RepFiles.RepWgts_May19;
Run;

/* Sort self-response replicate weights */
Proc Sort Data=repjul18;
  By QstNum OccurNum;
Run;

Proc Sort Data=repjan19;
  By QstNum OccurNum;
Run;

Proc Sort Data=repmay19;
  By QstNum OccurNum;
Run;

/* Merge survey and replicate weight data */
Data cpsjul18;
  Merge cpsjul18(In=In1)
        repjul18(In=In2);
  By QstNum OccurNum;
  If In1;
  If Not In2 Then Put "Error in Merge: " QstNum= OccurNum=;
Run;

Data cpsjan19;
  Merge cpsjan19(In=In1)
        repjan19(In=In2);
  By QstNum OccurNum;
  If In1;
  If Not In2 Then Put "Error in Merge: " QstNum= OccurNum=;
Run;
```


```
Data cpsmay19;
  Merge cpsmay19(In=In1)
        repmay19(In=In2);
  By QstNum OccurNum;
  If In1;
  If Not In2 Then Put "Error in Merge: " QstNum= OccurNum=;
Run;

/* Concatenate the three datasets */
Data CPS1819;
  Set cpsjul18
      cpsjan19
        cpsmay19;
Run;

/* Divide weights by number of months (surveys) being combined */
/* PWSRWgt is the weight from the main survey file.             */
/* RepWt0 is the same as PWSRWgt                                */
/* RepWgt1-RepWgt160 are the replicate weights                  */
Data CPS1819;
  Set CPS1819;
  Array Wgts(162) PWSRWgt RepWgt0-RepWgt160;
  Do I = 1 to 162;
    Wgts(I)=Wgts(I)/3;
      End;
Run;


/* Create recode variables for analysis */
Data CPS1819;
  Set CPS1819;
  /* CurrSmk: Current Cigarette Smoking Status */
  If PEA1=1 & PEA3 In (1,2) Then CurrSmk=1; /* Current Cigarette Smoker */
  Else If PEA1=1 & PEA3=3 Then CurrSmk=0;   /* Former Cigarette Smoker */
  Else If PEA1=2 Then CurrSmk=0;            /* Never Cigarette Smoker */
  Else CurrSmk=.;

  /*CurrVape: Current E-Cigarette Use*/
  If PEJ1A3_5=1 & PEJ2A3_5 in (1,2) then CurrVape=1;  /*Current*/
  Else If PEJ1A3_5=1 & PEJ2A3_5=3 then CurrVape=0;          /*Ever used, not currently
vaping*/
  Else if PEJ1A3_5=2 Then CurrVape=0;     /*Never vaped*/
  Else CurrVape=.;

  /* CigPD: Number of Cigarettes Per Day */
  If PEA1=1 & PEA3=1 & (1<=PTB1<=99) Then CigPD=PTB1;      /* Daily Smokers */
  Else If PEA1=1 & PEA3=2 & (1<=PEC1<=30) & (1<=PTC1A<=99) Then CigPD=(PEC1/30)*PTC1A; /*
Non-Daily Smokers */
  Else CigPD=.;

  /* Sex: Sex */   /* <= THIS ISN'T NECESSARY, COULD JUST USE PESEX */
  If PESex=1 Then Sex=1;        /* Male */
  Else If PESex=2 Then Sex=0;   /* Female */
```


```
  /* AgeGrp: Age Group */
  If (18<=PrtAge<=24) Then AgeGrp=1;        /* 18-24 */
  Else If (25<=PrtAge<=34) Then AgeGrp=2;   /* 25-34 */
  Else If (35<=PrtAge<=44) Then AgeGrp=3;   /* 35-44 */
  Else If (45<=PrtAge<=54) Then AgeGrp=4;   /* 45-54 */
  Else If (55<=PrtAge<=64) Then AgeGrp=5;   /* 55-64 */
  Else If (PrtAge=>65) Then AgeGrp=6;       /* 65+ */
  Else AgeGrp=.;

  /* RaceEthn: Race/Ethnicity */
  If PEHSPNON=1 Then RaceEthn=3;     /* Hispanic */
  Else If PEHSPNON=2 & PTDTRace=1 Then RaceEthn=1;  /* Non-Hispanic, White Alone */
  Else If PEHSPNON=2 & PTDTRace=2 Then RaceEthn=2;  /* Non-Hispanic, Black Alone */
  Else RaceEthn=4;  /* Non-Hispanic, All other (including multiracial) */

  Label CurrSmk  = "Current Cigarette Smoking Status"
           CurrVape = "Current E-Cigarette Use"
           CigPD    = "Number of Cigarettes Per Day"
           AgeGrp   = "Ages"
           RaceEthn = "Race/Ethnicity";
  Keep Intrview PRS64 CurrSmk CurrVape CigPD Sex AgeGrp RaceEthn PWSRWgt RepWgt1-
RepWgt160;
  Format CurrSmk CurrSmkF. CurrVape CurrVapeF. Sex SexF. AgeGrp AgeGrpF. RaceEthn
RaceEthF.;
Run;

/* SAS Code for Generating Estimates for Table 1 */
Proc SurveyFreq Data=CPS1819 VarMethod=BRR (Fay=0.5);
  Tables CurrSmk / cl;
  Tables (Sex AgeGrp)*CurrSmk / cl row;
  Weight PWSRWgt;
  RepWeights RepWgt1-RepWgt160;
Run;

/* SUDAAN Code for Generating Estimates for Table 1 */

Proc Crosstab Data=CPS1819 Design=BRR;
  Weight PWSRWgt;
  RepWgt RepWgt1-RepWgt160/ADJFay=4;
  Class CurrSmk Sex AgeGrp/NoFreqs;
  Tables (Sex AgeGrp)*CurrSmk;
  Print NSum="Sample Size"
        WSum="Weighted Size"
        RowPer="Percent"
        SERow="Std Error"
        LowRow="Lower 95% CI"
        UpRow="Upper 95% CI"
        /Style=NCHS NSumFmt=F8.0 WSumFmt=F10.0 RowPerFmt=F9.4 SERowFmt=F7.4
        LowRowFmt=F9.4 UpRowFmt=F9.4;
Run;
```


```
/* SAS Code for Generating Estimates for Table 2 */
Proc SurveyMeans Data=CPS1819 VarMethod=BRR (Fay=0.5);
  Var CigPD;
  Domain CurrSmk;
  Weight PWSRWgt;
  RepWeights RepWgt1-RepWgt160;
Run;

Proc SurveyMeans Data=CPS1819 VarMethod=BRR (Fay=0.5);
  Var CigPD;
  Domain CurrSmk*Sex;
  Weight PWSRWgt;
  RepWeights RepWgt1-RepWgt160;
Run;

/*SUDAAN Code for Generating Estimates for Table 2*/
Proc Descript Data=CPS1819 Design=BRR;
  Weight PWSRWgt;
  RepWgt RepWgt1-RepWgt160/ADJFay=4;
  class Sex/noFreq;
  var CigPD;

  SubPopN CurrSmk=1/Name="Current Cigarette Smokers";
  Print NSum="Sample Size"
        WSum="Weighted Size"
        Mean="Mean"
        SEMean="Std Error"
        LowMean="Lower 95% CI"
        UpMean="Upper 95% CI"
        /Style=NCHS NSumFmt=F8.0 WSumFmt=F10.0 MeanFmt=F9.4 SEMeanFmt=F7.4
        LowMeanFmt=F9.4 UpMeanFmt=F9.4;
run;


/* SAS Code for Generating Estimates for Table 3 */
Proc SurveyLogistic Data=CPS1819 VarMethod=brr (Fay=0.5);
  Class Sex (ref='Female') AgeGrp (ref='18-24') RaceEthn (ref='Non-Hispanic, White
Alone') CurrSmk (ref='Not Current') / Param=Ref;
  Model CurrVape = Sex AgeGrp RaceEthn Currsmk;
  Weight PWSRWgt;
  Repweights RepWgt1-RepWgt160;
Run;


/* SUDAAN Code for Generating Estimates for Table 3 */
Proc RLogist Data=CPS1819 Design=BRR;
  Weight PWSRWgt;
  RepWgt RepWgt1-RepWgt160/ADJFay=4;
  Class Sex AgeGrp RaceEthn CurrSmk/NoFreq;
  Reflevel Sex=0 AgeGrp=0 RaceEthn=1 CurrSmk=0;
  Model CurrVape = Sex AgeGrp RaceEthn CurrSmk;
Run;
```