Tobacco Use Supplement
Current Population Survey

TUS-CPS HARMONIZED DATASET – 1992–2023 USER GUIDE

May 2025

Table of Contents

Page

Suggested citation for the Harmonized Dataset:

National Cancer Institute. 2025. Tobacco Use Supplement to the Current Population Survey Harmonized Data, 1992-2023. cancercontrol.cancer.gov/tus-cps.

To cite the Technical Documentation and Data Files, replace the URL in the above citation with:

cancercontrol.cancer.gov/brp/tcrb/tus-cps/results.

Introduction

Overview

Staff from the National Cancer Institute (NCI) have harmonized Tobacco Use Supplement to the Current Population Survey (TUS-CPS) data from 1992 through 2023 to provide a more robust data source, increasing sample size and allowing for more seamless tracking of trends over time. The purpose of this user guide is to provide background on the design of the TUS-CPS Harmonized Dataset and guidance on using the Dataset to conduct weighted analyses. Specifically, this document provides guidance on conducting analyses using harmonized data while incorporating self-response and replicate weights.

Sections 1 and 2 of this document provide overviews of the CPS and TUS, respectively. Section 3 describes the data harmonization process and the various data files available for download. Section 4 details how replicate weights were generated. Section 5 contains examples of analyses using the harmonized data and replicate weights, including statistical code for each example in SAS.

Information included and examples shown in this document are not exhaustive; rather, they are intended to serve as an orientation to working with the TUS-CPS Harmonized Dataset. For greater detail, see "More Information" below and the various links included in each section of this document.

Summary of Files

All files associated with the Harmonized Dataset can be found on the TUS-CPS webpage at cancercontrol.cancer.gov/brp/tcrb/tus-cps/results.

- **Harmonized Dataset and Read-In Programs**

  Under the Harmonized Dataset section of the TUS-CPS Results page, there is a zip file containing the 1992–2023 Harmonized Data file (in .dat form), an accompanying read-in program, and a SAS format file.
  cancercontrol.cancer.gov/brp/tcrb/tus-cps/results

- **Replicate Weight Files and Read-In Programs**

  Under the Replicate Weights section of the TUS-CPS Results page, there are three separate zip files containing replicate weights for calculating variance estimates. The three replicate weight zip files, corresponding to the 1992-1993, 1995–2003, and 2006–2023 survey waves, account for the increasing number of replicate weights over time due to major redesigns to the CPS. This section of the webpage also contains programs to read in the replicate weights and merge them with the survey data.
  cancercontrol.cancer.gov/brp/tcrb/tus-cps/results

- **Harmonized Dataset Variable Crosswalk**

  Under the Harmonized Dataset section of the TUS-CPS Results page, the crosswalk can be used to identify variables that were harmonized (i.e., matched or made to match) across survey waves. Some variables contained in this file include notes explaining how the harmonized variable was constructed.
  cancercontrol.cancer.gov/sites/default/files/2025-06/HarmTUSCPS_1992-2023Crosswalk_508c.pdf

- **Harmonized Data Dictionary**

  Under the Harmonized Dataset section of the TUS-CPS Results page, the data dictionary provides details on harmonized variables and their response categories.
  cancercontrol.cancer.gov/sites/default/files/2025-06/HarmTUSCPS_1992-2023DataDict_508c.pdf

- **SAS Contents of Harmonized Database**

  Under the Harmonized Dataset section of the TUS-CPS Results page, the SAS output from a PROC CONTENTS of the entire TUS-CPS Harmonized Dataset is provided.
  cancercontrol.cancer.gov/sites/default/files/2025-06/HarmTUSCPS_1992-2023Contents_508c.pdf

- **Descriptive Frequency Tables of Harmonized Variables**

  Under the Harmonized Dataset section of the TUS-CPS Results page, this file contains frequency tables for all harmonized variables by survey wave.
  cancercontrol.cancer.gov/sites/default/files/2025-06/HarmTUSCPS_1992-2023Tables_508c.pdf

More Information

Information and examples included in this user guide are not exhaustive; rather, they are intended to serve as an orientation to working with the TUS-CPS Harmonized Dataset. Below are links to various webpages and documents that supplement the information presented here; these links also appear throughout the rest of the document.

Tobacco Use Supplement

- **TUS-CPS Homepage**

  Provides information about past, current, and future supplements, data reports, fact sheets, and resources for using TUS-CPS data.
  cancercontrol.cancer.gov/tus-cps

- **TUS-CPS Results**

  Connects users with questionnaires, data files, reports, replicate weight files, technical documentation, and user guides for all individual survey waves and the Harmonized Dataset.
  cancercontrol.cancer.gov/brp/tcrb/tus-cps/results

- **TUS-CPS Linkages**

  Provides information on other CPS supplements and studies (e.g., Tobacco Longitudinal Mortality Study) to which TUS-CPS data may be linked for additional analyses.
  cancercontrol.cancer.gov/brp/tcrb/tus-cps/linkages

- **User Workshops & Webinars**

  Provides links to archived recordings and materials from user workshops and webinars, including content on conducting weighted analyses, using small area estimation techniques, and other topics.
  cancercontrol.cancer.gov/brp/tcrb/tus-cps/workshops

Current Population Survey

- **Current Population Survey Homepage**

  census.gov/programs-surveys/cps.html

- **CPS Complete Technical Documentation**

  census.gov/programs-surveys/cps/technical-documentation/complete.html

- **Technical Paper 77**

  Provides greater detail about the design and methodology of the CPS. This information can also be found on the CPS Complete Technical Documentation webpage.
  www2.census.gov/programs-surveys/cps/methodology/CPS-Tech-Paper-77.pdf

User Support

Questions about this user guide or about the TUS-CPS Harmonized Dataset should be directed to NCI's Division of Cancer Control and Population Sciences, Behavioral Research Program at **ncidccpsbrpadvances@mail.nih.gov**.

1 Current Population Survey

Background

The Current Population Survey (CPS) is the source of the official government statistics on employment and unemployment. The CPS has been conducted monthly for more than 60 years. Additionally, the CPS provides the only available distribution of workers by the number of hours worked (as distinguished from aggregate or average hours for an industry), permitting separate analyses of part-time workers, workers on overtime, etc. The CPS not only collects information on individuals currently in the labor force but also on those who are outside the labor force, thus, data on current desire for work, past work experience, and job-seeking intentions are also available.

Although the main purpose of the CPS is to collect information on employment, an important secondary purpose of the survey is to collect information on demographic characteristics of the U.S. population such as age, sex, race, marital status, educational attainment, family relationship, occupation, and employment by industry. The statistics resulting from these demographic questions serve to update statistics generated every 10 years through the decennial census. Government policymakers and legislators consider these statistics as important indicators of our nation's economic situation and use them to plan and evaluate many government programs.

Current Population Survey Homepage
census.gov/programs-surveys/cps.html

Census Publications
census.gov/library/publications.html

CPS Sample Design

The primary objective of the CPS is to produce national and state estimates of labor force characteristics of the civilian noninstitutional population (CNP) ages 16 and older. Currently, the U.S. Census Bureau obtains interviews from approximately 54,000 households monthly. These households are scientifically selected based on area of residence to represent the nation, individual states, and other specified areas. Each household is interviewed monthly for four consecutive months in one year, and again for the corresponding time period one year later. This technique enables the Census Bureau to obtain reliable month-to-month and year-to-year comparisons at a reasonable cost while minimizing the inconvenience to any one household.

The CPS sample is a probability sample based on a stratified two-stage sampling scheme: selection of sample primary sampling units (PSUs) and selection of sample housing units within those PSUs. In general, the CPS sample is selected from lists of addresses obtained from the Master Address File (MAF) with updates from the United

States Postal Service (USPS) twice a year. The MAF is the Census Bureau's permanent list of addresses, including their geographic locations, for individual living quarters. It is continuously maintained through partnerships with the USPS; federal, state, regional, and local agencies; and the private sector. Historically, the CPS sample has been redesigned after each decennial census. Since 2015, the CPS sample has been based on information from the 2010 decennial census.

Approximately 72,000 sampled housing units are assigned for interview each month; about 60,000 are occupied and thus eligible for interview. The remaining units are determined to be destroyed, vacant, converted to nonresidential use, occupied by individuals usually residing elsewhere, or ineligible for interview for other reasons. Approximately 10 percent of the 60,000 occupied housing units are not interviewed in a given month for various reasons (e.g., temporary absence due to vacation, unable to contact residents after repeated attempts, inability of contacts to respond, or refusals to cooperate). The resulting 54,000 interviewed households contain approximately 108,000 people 15 years and older, approximately 27,000 children 0–14 years old, and about 450 Armed Forces members living with civilians either on or off base.

Geographic Limitations

The CPS sample was selected so that specific national and state-level statistical reliability criteria would be met. However, estimates calculated for geographic areas identified on the microdata file that are smaller than states are not as reliable.

CPS Complete Technical Documentation
census.gov/programs-surveys/cps/technical-documentation/complete.html

Technical Paper 77
www2.census.gov/programs-surveys/cps/methodology/CPS-Tech-Paper-77.pdf

2 Tobacco Use Supplement

Background

As discussed in Section 1, the CPS also collects information on topics other than employment. This information is collected periodically in the form of supplements to the basic questionnaire. The TUS-CPS is one such supplement.

The TUS-CPS is the largest nationally representative survey of adult tobacco use in the United States, with about 115,000–230,000 self-respondents per wave (the exact number of self-respondents varies by wave). As such, the survey is a key source of national, state, and sub-state (in larger metropolitan areas) data on tobacco use behaviors, attitudes, and policies in the United States. NCI has sponsored the TUS-CPS every three to four years since the survey's inception in 1992, with co-sponsorship from the U.S. Food and Drug Administration since 2014 and the Centers for Disease Control and Prevention from 2001 to 2007.

TUS-CPS data can be used by researchers to monitor tobacco control progress, conduct tobacco-related research, and evaluate tobacco control programs. Although the TUS-CPS has changed slightly between 1992 and 2023, it has generally covered topics such as current cigarette smoking status and smoking history, quit attempts and intention to quit, workplace smoking restrictions, attitudes toward smoke-free policies, and use of non-cigarette tobacco products.

Data Collection

To date, the TUS-CPS was fielded in 1992-1993, 1995-1996, 1998-1999, 2000, 2001-2002, 2003, 2006-2007, 2010-2011, 2014-2015, 2018-2019, and 2022-2023. Within each survey wave, data are collected at similar timepoints within three separate months (except for the year 2000, where data were only collected for two months).

Survey Wave	Data Collection Months Within Each Wave
1992-1993	September 1992, January 1993, May 1993
1995-1996	September 1995, January 1996, May 1996
1998-1999	September 1998, January 1999, May 1999
2000	January 2000, May 2000
2001-2002	June 2001, November 2001, February 2002
2003	February 2003, June 2003, November 2003
2006-2007	May 2006, August 2006, January 2007
2010-2011	May 2010, August 2010, January 2011
2011	May 2011
2014-2015	July 2014, January 2015, May 2015
2018-2019	July 2018, January 2019, May 2019
2022-2023	September 2022, January 2023, May 2023

Respondents are typically included in a TUS-CPS survey wave only once. Data collection months within a survey wave are generally spaced far enough apart to allow for complete turnover of CPS panels. As such, very few survey waves contain individuals with multiple self-responses. Exceptions include May 2010-2011 (a special, longitudinal TUS-CPS cohort), February 2002-2003, January and May 1999, and January and May 2000.

TUS-CPS Sample Design

As a supplement to the CPS, the TUS-CPS uses the sampling methodology of its parent survey, with some additional criteria, which are described below. In addition, some aspects of TUS-CPS sampling methodology have changed over the years. For details about specific survey waves, please refer to the appropriate technical documentation for that wave.

TUS-CPS Results
cancercontrol.cancer.gov/brp/tcrb/tus-cps/results

Currently, individuals ages 18 and older who have completed the CPS core survey are eligible for participation in the TUS-CPS. However, prior to January 2007, the TUS also included individuals ages 15–17. For consistency, all survey waves of the Harmonized Dataset only include respondents ages 18 and older.

Beginning in 2015, to decrease household response burden, the number of self-response interviews was limited in larger households. 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 for self-interview. Although proxy responses are collected when an individual selected for self-interview cannot be reached after a specified number of attempts, the Harmonized Dataset only contains data from self-respondents.

3 TUS-CPS Harmonized Data, 1992–2023

Background

Staff from the NCI TUS-CPS team harmonized the data from the 1992–2023 waves of the TUS-CPS to provide a more robust data source. In addition to tobacco-related measures, select variables from the CPS core survey were also harmonized, providing additional sociodemographic and employment data. In total, the Harmonized Dataset includes ten waves of data:

Survey Wave	Data Collection Months Within Each Wave
1992-1993	September 1992, January 1993, May 1993
1995-1996	September 1995, January 1996, May 1996
1998-1999	September 1998, January 1999, May 1999
2000	January 2000, May 2000
2001-2002	June 2001, November 2001, February 2002
2003	February 2003, June 2003, November 2003
2006-2007	May 2006, August 2006, January 2007
2010-2011	May 2010, August 2010, January 2011
2014-2015	July 2014, January 2015, May 2015
2018-2019	July 2018, January 2019, May 2019
2022-2023	September 2022, January 2023, May 2023

Initial Data Harmonization (1992–2015)

TUS-CPS data were initially harmonized to cover select TUS and CPS core variables with similar content from 1992 through 2015. Items that were used in two or more waves were considered eligible for harmonization. The initial harmonization process involved reviewing each eligible variable across the available survey waves and assigning them to one of the following categories:

- Variables to harmonize “as is” remained consistent over time and were harmonized in their current state.
- Variables to harmonize after adjustments had slight wording changes over time or slight variations in the universe of respondents across survey waves. These items were first adjusted to make the resultant data more complete. All adjustments were documented, and this information can be found in the notes section of the Harmonized Dataset Variable Crosswalk.
- Variables to drop had significant variations in wording, structure, or universe and were not harmonized.

Variables were not harmonized if there were fewer than two waves of data available, the question wording and/or structure was too inconsistent over time, or there were substantial changes in the universe of respondents over time.

Further, only self-reported responses were included in the Harmonized Dataset, as proxy data are limited and may present issues with validity. The Harmonized Dataset also only includes data from respondents ages 18 and older.

Harmonized variables were assigned one variable name across all waves, and a flag variable ("SURWAVE") was created to keep track of the survey year.

Updates to the Harmonized Dataset (1992–2023)

Since the initial harmonization of the TUS-CPS, the file has been updated twice, first to add data from the 2018-2019 wave, and then to add data from the 2022-2023 wave. Both updates involved two actions:

1. Adding newer data for previously harmonized variables, where possible. This involved adding data from self-respondents in the newest wave of the TUS-CPS for those variables which previously existed in the Harmonized Dataset. Variables that had been dropped from the most recent survey wave or that had significant changes to their question wording, structure, or respondent universe did not have new data added to the Harmonized Dataset.

2. Adding data for new variables (i.e., variables not included in the previous version of the Harmonized Dataset) for existing and new self-respondents. These variables could have been newly eligible for harmonization because (a) they now had two or more waves of data, or (b) they were determined to be of interest for harmonization by the NCI TUS-CPS team. New variables were harmonized following the same approach used for the initial harmonization process as described above. For a listing of new variables from the most recent update, please see Appendix 1.

Harmonized Data Files

All files associated with the Harmonized Dataset can be found on the TUS-CPS webpage at cancercontrol.cancer.gov/brp/tcrb/tus-cps/results.

- **Harmonized Dataset and Read-In Programs**

  Under the Harmonized Dataset section of the TUS-CPS Results page, there is a zip file containing the 1992–2023 Harmonized Data file (in .dat form), an accompanying read-in program, and a SAS format file.
  cancercontrol.cancer.gov/brp/tcrb/tus-cps/results

- **Replicate Weight Files and Read-In Programs**

  Under the Replicate Weights section of the TUS-CPS Results page, there are three separate zip files containing replicate weights for calculating variance estimates. The three replicate weight zip files correspond to the 1992-1993, 1995–2003, and 2006–

2023 survey waves, to account for the increasing number of replicate weights over time due to major redesigns to the CPS. This section of the webpage also contains programs to read in the replicate weights and merge them with the survey data. cancercontrol.cancer.gov/brp/tcrb/tus-cps/results

- **Harmonized Dataset Variable Crosswalk**

  Under the Harmonized Dataset section of the TUS-CPS Results page, the crosswalk can be used to identify variables that were harmonized (i.e., matched or made to match) across survey waves. Some variables contained in this file include notes explaining how the harmonized variable was constructed. cancercontrol.cancer.gov/sites/default/files/2025-06/HarmTUSCPS_1992-2023Crosswalk_508c.pdf

- **Harmonized Data Dictionary**

  Under the Harmonized Dataset section of the TUS-CPS Results page, the data dictionary provides details on harmonized variables and their response categories. cancercontrol.cancer.gov/sites/default/files/2025-06/HarmTUSCPS_1992-2023DataDict_508c.pdf

- **SAS Contents of Harmonized Database**

  Under the Harmonized Dataset section of the TUS-CPS Results page, the SAS output from a PROC CONTENTS of the entire TUS-CPS Harmonized Dataset is provided. cancercontrol.cancer.gov/sites/default/files/2025-06/HarmTUSCPS_1992-2023Contents_508c.pdf

- **Descriptive Frequency Tables of Harmonized Variables**

  Under the Harmonized Dataset section of the TUS-CPS Results page, this file contains frequency tables for all harmonized variables by survey wave. cancercontrol.cancer.gov/sites/default/files/2025-06/HarmTUSCPS_1992-2023Tables_508c.pdf

4 TUS-CPS Weighted Data

Background

The Harmonized TUS-CPS Dataset includes self-response full-sample survey weights and replicate weights, which are provided in separate files.

In each TUS-CPS data wave, 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, a series of weight variables, contain the information necessary for correctly computing the standard errors of point estimates, accounting for the complex survey design. Although individual survey waves of TUS-CPS include both self-response and self plus proxy full-sample survey weights and replicate weights, the Harmonized Dataset includes full-sample survey weights and replicate weights for self-respondents only.

The process of creating the TUS survey weights begins with the final CPS weights, which incorporate the probability of selection into CPS, adjust for non-response to CPS, and calibrate to the U.S. noninstitutionalized population. For this reason, we will briefly describe the creation of CPS weights first, and then the TUS weights.

CPS Weights

First, base weights are assigned for each CPS sample unit (individuals), which represent the inverse of a unit's probability of selection. Because most sample units within a state have the same probability of selection, they have the same base weight. Base weights are then adjusted for nonresponse to reduce bias that would arise from ignoring housing units that do not respond. There are two main types of nonresponse. The first is item nonresponse, which arises when a cooperating housing unit fails or refuses to provide some specific items of information. The second is unit nonresponse, which arises when the field representative is unable to collect any survey data from an eligible sample housing unit. This nonresponse adjustment is made within clusters of similar sample areas that are usually, but not necessarily, contained within a state. At this point, records for all individuals in the same household have the same weight because the nonresponse-adjusted weights depend only on household characteristics.

After base weights are adjusted for nonresponse, the ratio estimation procedure is applied. The distribution of demographic characteristics (e.g., age, race, sex, and state of residence) in the CPS sample in any given month may differ somewhat from that of the true population. Because these characteristics are closely correlated with labor force participation and other primary measurements estimated from the sample, the variance of sample estimates based on these characteristics can be reduced when the sample population distribution is brought as closely into agreement as possible with the known distribution of the entire population with respect to these characteristics. This is accomplished by adjusting the weights through a series of benchmarking adjustments as follows:

First-stage ratio adjustment reduces the contribution to variance that results from selecting a sample of PSUs rather than using all PSUs in the nation (between-PSU variance). Nonresponse-adjusted weights are post-stratified to match the state-specific Black alone/non-Black alone population distribution from independent population controls, then to match the national demographic totals for age by sex by race by ethnicity, and, finally, to match the state-specific population totals for the age by sex by race groups, which are derived from the Census Bureau's Population Estimates Program.

Second-stage ratio adjustment further reduces variances by benchmarking CPS estimates of the population to independent estimates of the current population. Weights are iteratively adjusted so that aggregated CPS sample estimates match independent estimates of population controls in various age/sex/race and age/sex/ethnicity cells at the national level. The following three sets of controls are used:

1. State/sex/age: the civilian noninstitutional population for the states by sex and age in years (0–15, 16–44, 45+)

2. Ethnicity/sex/age: total national civilian noninstitutional population for 36 Hispanic and non-Hispanic age/sex cells

3. Race/sex/age: total national civilian noninstitutional population for 56 White, 36 Black, and 26 "residual race" age/sex cells

For more information on CPS sampling and weights:

CPS Complete Technical Documentation
census.gov/programs-surveys/cps/technical-documentation/complete.html

Technical Paper 77
www2.census.gov/programs-surveys/cps/methodology/CPS-Tech-Paper-77.pdf

TUS-CPS Survey Weights

To create the TUS full-sample self-response survey weights, CPS weights are adjusted to account for supplement self-nonresponse (e.g., a CPS respondent selected for a self-response TUS interview declines). These weights are post-stratified to match the state-specific Black alone/non-Black alone population distribution from the corresponding decennial census used for the sample design, then to match the national demographic totals for age by sex by race by origin groups, and, finally, to match state-specific population totals for age by sex by race groups, which are derived from the Census Bureau's Population Estimates Program.

TUS-CPS Replicate Weights

TUS-CPS replicate weights are generated 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 the formulae used to measure variation across the replicated estimates. Methods include 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 included in Section 5. 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, for example, 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:

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

As described above, the TUS-CPS Harmonized Dataset includes TUS self-response full-sample survey and replicate weights from each individual survey wave. The number of replicate weights used for TUS-CPS has changed over time, and, accordingly, the replicate weights are provided in three separate files: years 1992-1993, years 1995–2003, and years 2006–2023. To create these files, replicate weights from years/months with same number of replicate weights were combined/stacked together, and year/month were included in the file as a source indicator. A separate file with

supplemental SAS code for reading in the data is also available for download. Appropriate replicate weights need to be merged with the main harmonized dataset before conducting analyses. See Section 5 for some example instructions.

Survey Wave	Number of Replicate Weights
1992-1993	48
1995-1996	80
1998-1999	80
2000	80
2001-2002	80
2003	80
2006-2007	160
2010-2011	160
2014-2015	160
2018-2019	160
2022-2023	160

For replicate weight files by wave, see**:**

TUS-CPS Results
cancercontrol.cancer.gov/brp/tcrb/tus-cps/results

Additional resource on replicate weights and variance estimation:

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

5 Analyzing TUS-CPS Harmonized Data: Examples in SAS

Introduction

The following examples illustrate three approaches to using TUS-CPS harmonized data: using only one wave of data (Example 1) and analyzing multiple waves of data (Examples 2a and 2b). Each example discusses related considerations, includes corresponding SAS code, and shows statistical output.

Preparing the Harmonized Data

Before conducting specific analyses, the main harmonized survey data file must be merged with the harmonized replicate weight files. The following instructions describe how to merge the file containing all waves of harmonized data with all harmonized replicate weight files to create one dataset, which can be used as a basis for the examples. However, the resulting dataset is quite large. In practice, researchers can improve run times by creating a smaller dataset, which involves taking a subset of the harmonized survey data and merging it with replicate weights for the corresponding waves only.

All data files, as well as SAS programs to read in the data files, can be downloaded from the TUS-CPS website. Refer to the Introduction to this user guide or Section 3 for a description of where to locate these files. Save all files in the same location on your computer, to simplify the process of importing files into SAS.

Filename	Description
Harmonzd.tus_cps.1992.through.2023.formats.sas	SAS program to input variable formats
Harmonzd.tus_cps.1992.through.2023.sas	SAS program to read in the main survey file
Harmonzd.tus_cps.1992.through.2023.dat	Main survey data file
harmonzd.tus_cps.1992.through.2023.add.replicate.weights.sas	SAS program to read in the replicate weight files and merge them with the survey data
Harmonzd.tus_cps.1992.through.2023.replicate.wgts.06_23.dat	Replicate weights 2006–2023 data file
Harmonzd.tus_cps.1992.through.2023.replicate.wgts.92_93.dat	Replicate weights 1992-1993 data file
Harmonzd.tus_cps.1992.through.2023.replicate.wgts.95_03.dat	Replicate weights 1995–2003 data file

1. Open the SAS program: Harmonzd.tus_cps.1992.through.2023.sas.

   Modify libraries and file paths such that they are associated with the folder containing survey data and replicate weight files downloaded from the TUS-CPS website, then run the program to read in the main survey data file. ** *Please note that this creates a new permanent dataset called "Harmon" and stores it in the library specified at the beginning of the program.* **

2. Open the SAS program: harmonzd.tus_cps.1992.through.2023.add.replicate.weights.sas.

   Modify libraries and file paths, then run the program to read in the replicate weight files and merge them with the survey data. This creates a new dataset, also called "Harmon," that is stored in the default temporary "Work" library in SAS.

   From the Log:

   ```
   NOTE: There were 1,845,090 observations read from the data set WORK.HARMON.
   NOTE: There were 1,845,090 observations read from the data set WORK.REPS.
   NOTE: The data set WORK.HARMON has 1,845,090 observations and 451 variables.
   ```

3. Create a permanent dataset called "Harmon2" and store it in the library "MyLib," specified in Step 1. ** *Please note that the size of the complete Harmonized Dataset, including survey data and replicate weights, is very large and may cause programs to run slowly. Users may wish to subset the data to specific variables of interest (see examples).* **

   ```
   Data MyLib.Harmon2;
     Set Harmon;
   Run;
   ```

Example 1. Using only one wave of TUS-CPS harmonized data

Objectives:

1. Calculate mean cigarettes per day among current smokers during the 2018-2019 wave, overall and by age group, and

2. Use ANOVA to test for differences in mean cigarettes per day among current smokers across age groups.

Some research questions may require that analyses are conducted within a particular subset of respondents. For example, when examining the mean number of cigarettes smoked per day among *current* smokers, respondents who are *former* or *never* smokers are not of interest. When working with complex survey data where strata and PSU information are used for variance estimation, it is important to keep all respondents in the dataset throughout analysis because subsetting may result in underestimation of the variance. For those surveys providing strata and PSU information, proper methods for conducting subgroup analyses incorporate sampling information for all observations, even those that are not in the subpopulation of interest. To conduct subgroup analyses using SAS SURVEY procedures, researchers should use DOMAIN statements and avoid use of WHERE statements. For more information, please see Paper 449-2013 from the SAS Global Forum 2013. However, given the way that variance is computed (see Box A in Section 4) in surveys that include replicate weights, such as the TUS-CPS, keeping all observations and using DOMAIN statements will not produce a different result than taking a subset of observations and using WHERE statements. For consistency, these examples use DOMAIN statements for subgroup analysis and, when analyzing a smaller set of survey waves from the Harmonized Dataset, use WHERE or IF statements to subset the data.

The following steps outline how to prepare the 2018-2019 data for calculating mean cigarettes smoked per day among current smokers overall and by age group. **This example uses the datafile "Harmon2," created in Steps 1–3 of the previous Preparing the Data section.**

1. Prepare to analyze the data by respecifying (if necessary) the library and file path, including the formats provided with the Harmonized Dataset, and creating any new format that will be used in the analysis. *** Please note that the library name "MyLib" and file path "T:\TUS\data\" are examples. ***

```
Libname MyLib "T:\TUS\data\";

%Include "harmonzd.tus_cps.1992.through.2023.formats.sas";

Proc Format;
  Value CurrSmkF
    0 = "Non-Smoker"
    1 = "Current Cigarette Smoker"
    ;
  Value Age_CatF
    1 = "18-24 years old"
```

```
    2 = "25-44 years old"
    3 = "45-64 years old"
    4 = "65 years or older"
    ;
```

2. Due to the large size of the harmonized data file ("Harmon2"), users may opt to create a new dataset (here called "Example1") that is a subset of the larger datafile to minimize run times. From the dataset "Harmon2", select the TUS-CPS self-respondents who completed the 2018-2019 survey wave using the variable "SURWAVE".

```
Data Example1;
  Set MyLib.Harmon2;
  If SURWAVE=10; /*2018-2019 Survey Wave*/
Run;
```

   From the log:

```
NOTE: There were 1,845,090 observations read from the data set MYLIB.HARMON2.
NOTE: The data set WORK.EXAMPLE1 has 137,471 observations and 451 variables.
```

3. Recode variables needed for analysis: Current Cigarette Smoking Status ("CurrSmk": yes/no), Cigarettes Smoked Per Day ("CigPD": continuous, among current smokers), and Age Group ("Age_Cat": 18 to 24, 25 to 44, 45 to 64, and 65 or older).

```
Data Example1;
  Set Example1;
  /* CurrSmk: Current Cigarette Smoking Status */
  If CigStat in (2,3) Then CurrSmk=1; /* Current Cigarette Smoker */
  Else If CigStat in (1,4) Then CurrSmk=0; /* Non-Smoker */
  Else CurrSmk=.;

  /* CigPD: Number of Cigarettes Per Day */
  If CigStat=2 & (0<=CPDD<=99) Then CigPD=CPDD; /* Daily Smokers */
  Else If CigStat=3 & (0<=CPDS<=30) Then CigPD=CPDS; /* Non-Daily
Smokers */
  Else CigPD=.;

  /* Age_Cat: Age Group*/
  If 18 <= Age <= 24 Then Age_Cat=1; /* 18-24 years old */
  Else If 25 <= Age <= 44 Then Age_Cat=2; /* 25-44 years old */
  Else if 45 <= Age <= 64 Then Age_Cat=3; /* 45-64 years old */
  Else if Age => 65 Then Age_Cat=4; /* 65 years or older */

  Label CurrSmk = "Current Cigarette Smoking Status"
        CigPD = "Number of Cigarettes Per Day"
        Age_Cat = "Age Group" ;

  Format CurrSmk CurrSmkF. Age_Cat Age_CatF.;

  Keep CurrSmk CigPD Age_Cat SmplWgt RepWt001-RepWt160;
Run;
```

4. Divide person-weights and replicate weights by 3 (because there are three months of data being combined for analysis: July 2018, January 2019, and May 2019). Otherwise, the dataset will yield a sample three times the size of the U.S. population. (Note: If using only the 2000 survey wave, person-weights and replicate weights should be divided by 2 because there are two months of data. Otherwise, the dataset will yield a sample two times the size of the U.S. population.)

```
/* SRWEIGHT is the weight from the main survey file.          */
/* SmplWgt is the same as SRWEIGHT                          */
/* RepWt001-RepWt160 are the replicate weights               */
Data Example1;
  Set Example1;
  Array Wgts(161) SmplWgt RepWt001-RepWt160;
  Do I = 1 to 161;
    Wgts(I)=Wgts(I)/3;
    End;

Run;
```

5. Calculate means and standard errors for cigarettes smoked per day by age group among current cigarette smoking (CURRSMK=1) adult self-respondents. Estimates will be weighted using the adjusted self-response person-weight (SMPLWGT) and self-response replicate weights (REPWT001-REPWT160). ** *Please note that earlier waves of data may have a different number of replicate weights. See Section 4 for a table of replicate weights included in each wave.* **

```
Proc SurveyMeans Data=Example1 VarMethod=BRR (Fay=0.5);
  Var CigPD;
  Domain CurrSmk;
  Weight SmplWgt;
  RepWeights RepWt001-RepWt160;
Run;

Proc SurveyMeans Data=Example1 VarMethod=BRR (Fay=0.5);
  Var CigPD;
  Domain CurrSmk*Age_Cat;
  Weight SmplWgt;
  RepWeights RepWt001-RepWt160;
Run;
```

6. Use ANOVA to test for differences in mean cigarettes smoked per day across age groups. ** *Please note that ANOVA does not use design information (i.e., replicate weights). A REPWEIGHTS statement is still included in the PROC SURVEYREG below because design information is used in estimating a regression model, but it is not necessary for this example.* **

```
Proc SurveyReg Data=Example1 VarMethod=BRR (Fay=0.5);
  Class Age_Cat (Ref="18-24 years old");
  Weight SmplWgt;
  RepWeights RepWt001-RepWt160;
  Domain CurrSmk ("Current Cigarette Smoker");
  Model CigPD = Age_Cat / anova;
Run;
```

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

Table 1. Unweighted and Weighted Mean Cigarettes Smoked Per Day Among Adult Smokers Overall and by Age Group, 2018-2019

	Unweighted	Weighted		
	N	Mean	95% CL	p-value
Overall	15,995	11.84	11.66–12.02	-
Age group				< 0.0001
18–24 years	695	9.16	8.45–9.86	
25–44 years	5,650	10.81	10.55–11.07	
45–64 years	6,925	13.10	12.84–13.36	
65+ years	2,725	12.37	11.99–12.74	

Source: TUS-CPS Harmonized Dataset, 2018-2019

Example 2. Analyzing two waves of TUS-CPS harmonized data

Objectives:

1. Calculate the unweighted frequencies and estimated weighted prevalence of current smoking overall and by sex during two consecutive waves, and

2. Estimate adjusted odds ratios and 95% confidence intervals for the association between sex and current smoking status.

Analyzing two consecutive waves of data at the same time can increase sample sizes and allow researchers to draw conclusions about a longer period of time. However, when working with TUS, it is important to pay attention to the number of replicate weights in the waves being pooled, as the number of replicate weights changes across waves. **Both parts of this example use the datafile "Harmon2," created in Steps 1–3 of the previous Preparing the Data section.**

Example 2a. Two waves with the same number of replicate weights (2006-2007 and 2010-2011)

1. Prepare to analyze the data by respecifying (if necessary) the library and file path, including the formats provided with the Harmonized Dataset, and creating any new format that will be used in the analysis. *** Please note that the library name "MyLib" and file path "T:\TUS\data\" are examples. ***

```
Libname MyLib "T:\TUS\data\";

%Include "harmonzd.tus_cps.1992.through.2023.formats.sas";

Proc Format;
  Value CurrSmkF
    0 = "Non-Smoker"
    1 = "Current Cigarette Smoker"
    ;
  Value MaleF
    0 = "Female"
    1 = "Male"
    ;
```

2. Due to the large size of the harmonized data file ("Harmon2"), users may opt to create a new dataset (here called "Example2a") that is a subset of the larger datafile to minimize run times. From the dataset "Harmon2", select the TUS-CPS self-respondents who completed the 2006-2007 and 2010-2011 survey waves using the variable "SURWAVE."

```
Data Example2a;
  Set MyLib.Harmon2;
  If SURWAVE in (7,8); /*2006-2007, 2010-2011 Survey Waves*/
Run;
```

From the log:

```
NOTE: There were 1,845,090 observations read from the data set MYLIB.HARMON2.
NOTE: The data set WORK.EXAMPLE2A has 343,388 observations and 451 variables.
```

3. Recode variables needed for analysis: Current Cigarette Smoking Status ("CurrSmk": yes/no) and Male Sex ("Male": yes/no).

```
Data Example2a;
  Set Example2a;

  /* CurrSmk: Current Cigarette Smoking Status */
  If CigStat in (2,3) Then CurrSmk=1; /* Current Cigarette Smoker */
    Else If CigStat in (1,4) Then CurrSmk=0; /* Non-Smoker */
    Else CurrSmk=.;

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

  Label CurrSmk = "Current Cigarette Smoking Status"
        Male = "Male Sex";
  Format CurrSmk CurrSmkF. Male MaleF.;
  Keep CurrSmk Male SmplWgt RepWt001-RepWt160;
Run;
```

4. Divide the person-weights and replicate weights by 6 (the number of months of data being combined for analysis), so that the weights total the average size of the U.S. population during the 2006-2007 and 2010-2011 periods.

```
Data Example2a;
  Set Example2a;
  Array Wgts(161) SmplWgt RepWt001-RepWt160;
  Do I = 1 to 161;
    Wgts(I)=Wgts(I)/6;
    End;
Run;
```

5. Calculate the estimated prevalence and 95% confidence intervals for current cigarette smoking. Estimates will be weighted using the adjusted self-response person-weight (SMPLWGT) and self-response replicate weights (REPWT001-REPWT160).

```
Proc SurveyFreq Data=Example2a VarMethod=BRR (Fay=0.5);
  Tables CurrSmk/CL;
  Tables Male*CurrSmk/Row CL;
  Weight SmplWgt;
  Repweights RepWt001-RepWt160;
Run;
```

6. Generate the odds ratio and 95% confidence interval for current cigarette smoking among males (vs. females). Estimates will be weighted using the adjusted self-response person-weight and self-response replicate weights.

```
Proc SurveyLogistic Data=Example2a VarMethod=BRR (Fay=0.5);
  Model CurrSmk (ref="Non-Smoker") = Male;
  Weight SmplWgt;
  RepWeights RepWt001-RepWt160;
Run;
```

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

Table 2a. Estimated Prevalence and Odds Ratio for Current Cigarette Smoking Among U.S. Adults by Sex, 2006-2011

	Unweighted	Weighted			
	N	N	Percent (95% CI)	Odds Ratio (95% CI)	p-value
Overall	59,112	38,678,139	17.26 (17.07–17.44)	-	-
Female	30,336	17,773,385	15.32 (15.11–15.53)	*ref*	*ref*
Male	28,776	20,904,755	19.34 (19.08–19.60)	1.33 (1.30–1.35)	< 0.0001

Source: TUS-CPS Harmonized Dataset, 2006-2011

Example 2b. Two waves with different numbers of replicate weights (2003 and 2006-2007)

In Example 2a, the two waves had the same number of replicate weights. In this example, the two waves being analyzed have different numbers of replicate weights: 2003 has 80 and 2006-2007 has 160. When combining survey waves that do not have the same number of replicate weights, a new set of replicate weights must be created, and the Fay's value used for analysis must change. To create the new set of replicate weights, the original replicate weights must be expanded and adjusted.

Expansion results in an even number of replicate weights for all observations, even if the survey waves being combined for analysis initially had different numbers of replicate weights. In this case, the 2003 wave uses 80 replicate weights and the 2006-2007 wave uses 160, so the expanded set will include 240 replicate weights (80 + 160) for all observations.

Adjustments are intended to 1) maintain the same variance for each survey wave before and after pooling, and 2) maintain a weighted sample size equal to the average size of the U.S. population during the 2003 and 2006-2007 periods. **For more information on generating replicate weights across survey waves with different numbers of replicates, see: surveillance.cancer.gov/reports/tech2020.01.pdf.**

The following steps outline how to prepare the 2003 and 2006-2007 data for calculating current smoking prevalence overall and by sex:

1. Prepare to analyze the data by respecifying (if necessary) the library and file path, including the formats provided with the Harmonized Dataset, and creating any new format that will be used in analysis. Formats included in this step are the same as in Example 2a. *** Please note that the library name "MyLib" and file path "T:\TUS\data\" are examples. ***

```
Libname MyLib "T:\TUS\data\";

%Include "harmonzd.tus_cps.1992.through.2023.formats.sas";

Proc Format;
  Value CurrSmkF
    0 = "Non-Smoker"
    1 = "Current Cigarette Smoker"
    ;
  Value MaleF
    0 = "Female"
    1 = "Male"
    ;
```

2. Due to the large size of the harmonized data file ("Harmon"), users may opt to create a new dataset (here called "Example2b") that is a subset of the larger datafile to minimize run times. From the dataset "Harmon2", select the TUS-CPS self-respondents who completed the 2003 and 2006-2007 survey waves using the variable "SURWAVE."

```
Data Example2b;
  Set MyLib.Harmon2;
  If SURWAVE in (6,7); /*2003, 2006-2007 Survey Waves*/
Run;
```

   From the log:

```
NOTE: There were 1,845,090 observations read from the data set MYLIB.HARMON2.
NOTE: The data set WORK.EXAMPLE2B has 355,833 observations and 451 variables.
```

3. Recode variables needed for analysis: Current Cigarette Smoking Status ("CurrSmk": yes/no) and Male Sex ("Male": yes/no).

```
Data Example2b;
  Set Example2b;

  /* CurrSmk: Current Cigarette Smoking Status */
  If CigStat in (2,3) Then CurrSmk=1; /* Current Cigarette Smoker */
  Else If CigStat in (1,4) Then CurrSmk=0; /* Non-Smoker */
  Else CurrSmk=.;

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

```
  Label CurrSmk = "Current Cigarette Smoking Status"
        Male = "Male Sex";
  Format CurrSmk CurrSmkF. Male MaleF.;
Run;
```

4. Divide the person-weights by 6 (the number of months of data being combined for analysis; three months from each survey wave). In the same DATA step, expand and adjust the replicate weights. Please note that the adjustment factor changes depending on the original number of replicate weights and the new/expanded number of replicate weights. See comments in the code for additional description of the expansion process and for specific adjustments.

```
Data Example2b;
  Set Example2b;
  Array OldR(160) RepWt001-RepWt160;
  Array NewR(240) NWgt001-NWgt240;

  NSmplWgt=SmplWgt/6;

            /*We need to create a new set of 240 adjusted replicate
            weights (80+160). For 2003, replicate weights 1 through 80
            will be the original 80 replicate weights divided by the
            number of surveys being analyzed (6) and adjusted to
            maintain the same variance. The remaining 160 replicate
            weights will be the individual's person-weight divided by
            the number of surveys.*/

  If SurWave=6 Then Do;
    Do I = 1 to 80;
      NewR(I)=(1/6) * (SmplWgt+(.866025*(OldR(I)-SmplWgt)));   /* .866025
= 1/2 x (Sqrt(240/80) */
    End;
    Do I = 81 to 240;
      NewR(I)=SRWEIGHT/6;
    End;
 End;

            /*We need to create a new set of 240 adjusted replicate
            weights (80+160). For 2006-07, the first 80 replicate
            weights will be the individual's person-weight divided by
            the number of surveys (6). Replicate weights 81 through 240
            will be the original 160 replicate weights divided by the
            number of surveys being analyzed and adjusted to maintain
            the same variance.*/

  Else Do;
    Do I = 1 to 80;
      NewR(I)=SRWEIGHT/6;
    End;
    Do I = 81 to 240;
      J = I - 80;
      NewR(I)=(1/6) * (SmplWgt+(.612372*(OldR(J)-SmplWgt)));  /* .612372
= 1/2 x (Sqrt(240/160)  */
    End;
  End;
```

```
Keep CurrSmk Male NSmplWgt NWgt001-NWgt240;
Run;
```

5. Estimate the prevalence for current cigarette smoking (CURRSMK) by sex among adult self-respondents. Estimates will be weighted using the adjusted self-response person-weight (NSMPLWGT), and the adjusted, expanded set of self-response replicate weights (NWGT001-NWGT240) will be used for variance estimation. Because the set of replicate weights was expanded, Fay's value is 0.75.

```
Proc SurveyFreq Data=Example2b VarMethod=BRR (Fay=0.75);
  Tables CurrSmk/Row CL;
  Tables Male*CurrSmk/Row CL;
  Weight NSmplWgt;
  Repweights NWgt001-NWgt240;
Run;
```

6. Generate the odds ratio and 95% confidence interval for current cigarette smoking among males (vs. females). Estimates will be weighted using the adjusted self-response person-weight and self-response replicate weights.

```
Proc SurveyLogistic Data=Example2b VarMethod=BRR (Fay=0.75);
  Model CurrSmk (ref="Non-Smoker") = Male;
  Weight NSmplWgt;
  RepWeights NWgt001-NWgt240;
Run;
```

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

Table 2b. Estimated Prevalence and Odds Ratio for Current Cigarette Smoking Among U.S. Adults by Sex, 2003–2007

	Unweighted	Weighted			
	N	N	Percent (95% CI)	Odds Ratio (95% CI)	p-value
Overall	66,145	40,286,815	18.69 (18.50–18.88)	-	-
Female	34,372	18,600,267	16.62 (16.41–16.84)	ref	ref
Male	31,773	21,686,548	20.91 (20.64–21.19)	1.33 (1.30–1.35)	< 0.0001

Source: TUS-CPS Harmonized Dataset, 2003–2007

Appendix 1: Variables Added in the 1992–2023 Harmonized Dataset

For more information on the construction of newly harmonized variables, please reference the Harmonized Dataset Variable Crosswalk and the Harmonized Data Dictionary (see links under Introduction and Section 3).

Variables available in two or more survey waves as of 2022-2023 that are newly harmonized include:

- Purchase of loose tobacco for “roll-your-own” cigarette, LOOSIES.
- Type of e-cigarette device, ECIGTYPE.
- Purchase own e-cigarette, BUYECIG.
- Purchase of e-cigarettes by box or pack, BUYPKECG.
- Price paid for last box or pack of e-cigarettes/pods/cartridges, PRCEBOX.
- Number of e-cigarettes/pods/cartridges in box or pack, NUMECIGS.
- Price paid for last single e-cigarette/pod/cartridge, PRCESNGL.
- Milliliters of e-liquid in a single bottle purchased, ECIGVOL.
- Indoor vaping or e-cigarette policy at place of work, WKPOLECG.
- Use of e-cigarettes or vaping in work area, WKSMK2EC.
- Indoor vaping or e-cigarette policy at home, HMECGPOL.

Variables determined to be of interest for harmonization by the NCI TUS-CPS team that are newly harmonized include:

- Three detailed Hispanic Origin Group variables, reflecting changes in response options over time:
  - 2014-2023, DTHISP14.
  - 2003-2011, DTHISP03.
  - 1992-2002, DTHISP92.
- A detailed Asian Race recoded variable for 2014-2023, DTASIAN.
- A Cohabitating Partner variable for 2010–2023, COHAB.

Newly harmonized variables have been added to the end of the file to enable existing variables to maintain a similar record position to prior years.