SAMHSA
Substance Abuse and Mental Health
Services Administration

2023 National Survey on Drug Use and Health

Public Use File Data Users’ Guide

2023 National Survey on Drug Use and Health: Public Use File Data Users' Guide

This publication was prepared for the Substance Abuse and Mental Health Services Administration (SAMHSA) under contract number 75S20322C00001 with SAMHSA, U.S. Department of Health and Human Services (HHS). Marlon Daniel served as government project officer and as contracting officer representative.

Recommended Citation

Center for Behavioral Health Statistics and Quality. (2025). *2023 National Survey on Drug Use and Health: Public use file data users' guide*. https://www.samhsa.gov/data/data-we-collect/nsduh/datafiles

Originating Office

Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, 5600 Fishers Lane, Rockville, MD 20857. Released 2025.

Electronic Access

Products may be downloaded at https://library.samhsa.gov/.

Disclaimer

Nothing in this document constitutes a direct or indirect endorsement by SAMHSA or HHS of any nonfederal entity's products, services, or policies.

Public Domain Notice

U.S. Department of Health and Human Services
Substance Abuse and Mental Health Services Administration
Center for Behavioral Health Statistics and Quality
Office of Population Surveys

Released 2025

Contents

Figures

Tables

Exhibits

1. Overview

This data users' guide describes how to analyze data using the public use file (PUF) for the National Survey on Drug Use and Health (NSDUH). It includes an overview of NSDUH history and trend breaks, as well as information on sample design, data collection, and response rates. This guide also describes how to use the questionnaire and codebook to identify analysis variables, how to use weighting variables, how to perform variance estimation, and how to conduct analyses with common statistical analysis software. NSDUH is sponsored by the Center for Behavioral Health Statistics and Quality (CBHSQ) within the Substance Abuse and Mental Health Services Administration (SAMHSA) and is conducted by RTI International.[1]

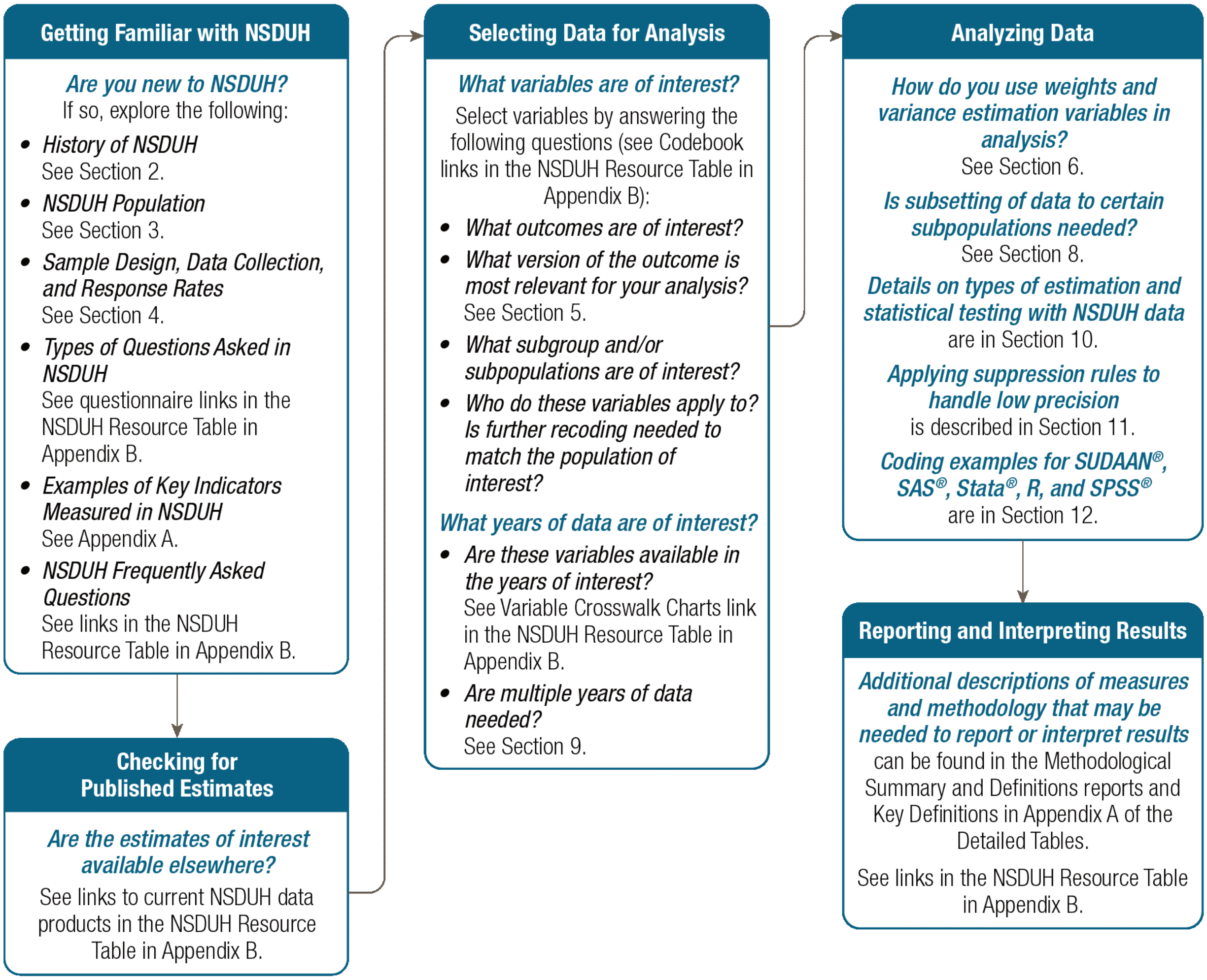
Statistical disclosure limitation procedures are implemented during the construction of the PUF to ensure any information that may be used to directly identify respondents is removed. These changes to the data may result in estimates that differ slightly from those created using the restricted-use file (RUF) and published in NSDUH data products.

Appendix A provides examples of key indicators measured in NSDUH. Links to additional NSDUH resources are included throughout the users' guide and in the NSDUH Resource Table in Appendix B.

Figure 1.1 presents a flowchart to guide data users on the process for conducting analyses using NSDUH data and where to find additional details on each step.

[1] RTI International is a trade name of Research Triangle Institute. RTI and the RTI logo are U.S. registered trademarks of Research Triangle Institute.

Figure 1.1 Getting Started with NSDUH Data



Getting Familiar with NSDUH

Are you new to NSDUH?
If so, explore the following:

- *History of NSDUH*
  See Section 2.
- *NSDUH Population*
  See Section 3.
- *Sample Design, Data Collection, and Response Rates*
  See Section 4.
- *Types of Questions Asked in NSDUH*
  See questionnaire links in the NSDUH Resource Table in Appendix B.
- *Examples of Key Indicators Measured in NSDUH*
  See Appendix A.
- *NSDUH Frequently Asked Questions*
  See links in the NSDUH Resource Table in Appendix B.

Checking for Published Estimates

Are the estimates of interest available elsewhere?
See links to current NSDUH data products in the NSDUH Resource Table in Appendix B.

Selecting Data for Analysis

What variables are of interest?
Select variables by answering the following questions (see Codebook links in the NSDUH Resource Table in Appendix B):

- *What outcomes are of interest?*
- *What version of the outcome is most relevant for your analysis?*
  See Section 5.
- *What subgroup and/or subpopulations are of interest?*
- *Who do these variables apply to? Is further recoding needed to match the population of interest?*

What years of data are of interest?

- *Are these variables available in the years of interest?*
  See Variable Crosswalk Charts link in the NSDUH Resource Table in Appendix B.
- *Are multiple years of data needed?*
  See Section 9.

Analyzing Data

How do you use weights and variance estimation variables in analysis?
See Section 6.

Is subsetting of data to certain subpopulations needed?
See Section 8.

Details on types of estimation and statistical testing with NSDUH data
are in Section 10.

Applying suppression rules to handle low precision
is described in Section 11.

Coding examples for SUDAAN®, SAS®, Stata®, R, and SPSS®
are in Section 12.

Reporting and Interpreting Results

Additional descriptions of measures and methodology that may be needed to report or interpret results
can be found in the Methodological Summary and Definitions reports and Key Definitions in Appendix A of the Detailed Tables.

See links in the NSDUH Resource Table in Appendix B.

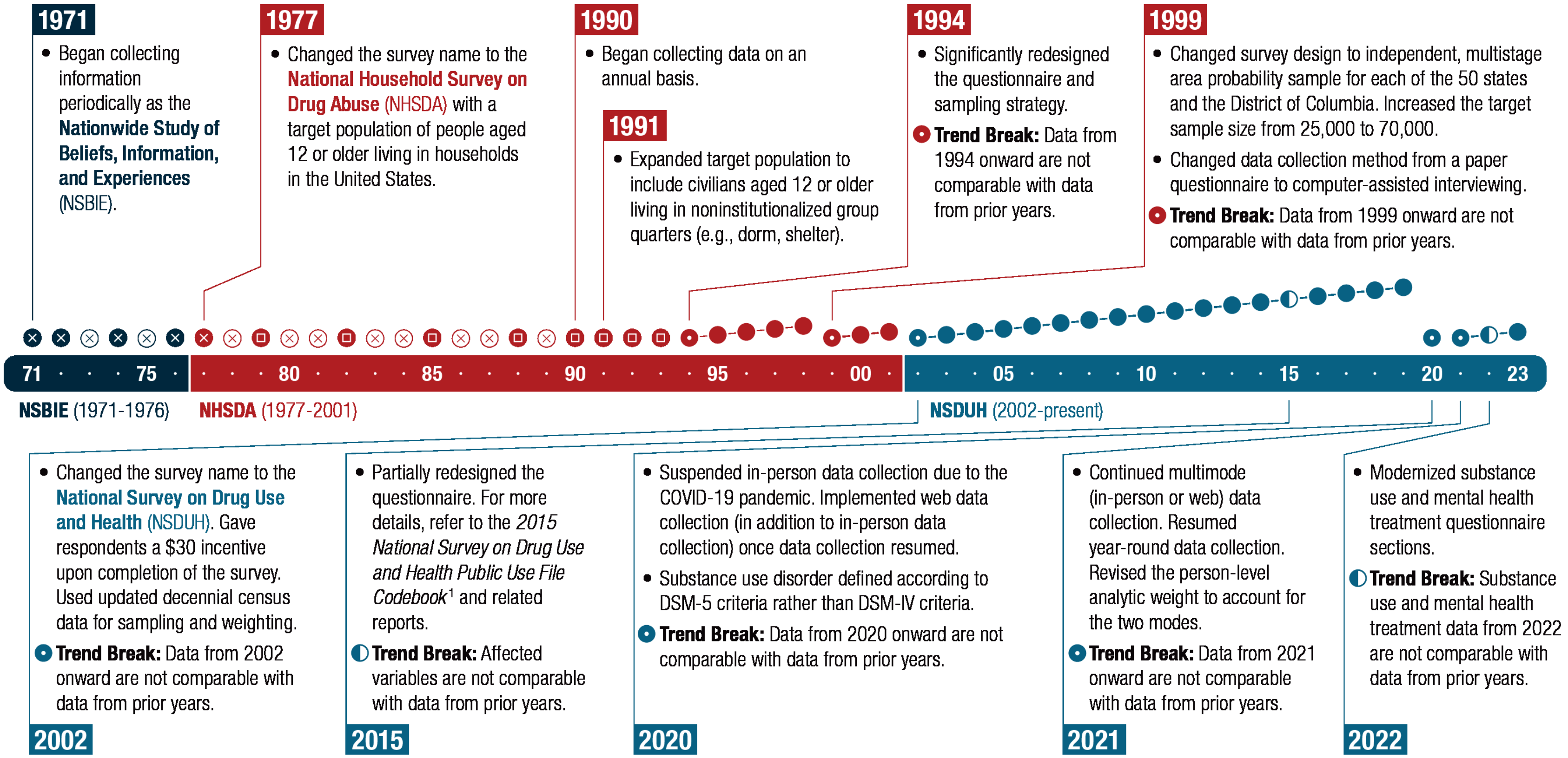
2. History of NSDUH and Trend Breaks

NSDUH is an annual survey of the civilian, noninstitutionalized population of the United States aged 12 years or older. It is the primary source of statistical information on the use of tobacco, alcohol, prescription psychotherapeutic drugs (pain relievers, tranquilizers, stimulants, and sedatives), and other illicit substances (e.g., marijuana, cocaine). The survey also collects information on substance use disorders, substance use treatment, mental health issues, and mental health treatment.

For a more detailed description of annual changes to the questionnaire, including updates to the measurement of individual questions or sections of the survey, refer to the year-specific Methodological Summary and Definitions at https://www.samhsa.gov/data/all-reports.

Figure 2.1 shows a timeline of survey changes and trend breaks that NSDUH has undergone over the years.

Figure 2.1 NSDUH Major Changes and Trend Breaks



1971

- Began collecting information periodically as the **Nationwide Study of Beliefs, Information, and Experiences** (NSBIE).

1977

- Changed the survey name to the **National Household Survey on Drug Abuse** (NHSDA) with a target population of people aged 12 or older living in households in the United States.

1990

- Began collecting data on an annual basis.

1991

- Expanded target population to include civilians aged 12 or older living in noninstitutionalized group quarters (e.g., dorm, shelter).

1994

- Significantly redesigned the questionnaire and sampling strategy.

⊙ **Trend Break:** Data from 1994 onward are not comparable with data from prior years.

1999

- Changed survey design to independent, multistage area probability sample for each of the 50 states and the District of Columbia. Increased the target sample size from 25,000 to 70,000.
- Changed data collection method from a paper questionnaire to computer-assisted interviewing.

⊙ **Trend Break:** Data from 1999 onward are not comparable with data from prior years.

NSBIE (1971-1976) **NHSDA** (1977-2001) **NSDUH** (2002-present)

2002

- Changed the survey name to the **National Survey on Drug Use and Health** (NSDUH). Gave respondents a $30 incentive upon completion of the survey. Used updated decennial census data for sampling and weighting.

⊙ **Trend Break:** Data from 2002 onward are not comparable with data from prior years.

2015

- Partially redesigned the questionnaire. For more details, refer to the *2015 National Survey on Drug Use and Health Public Use File Codebook*[1] and related reports.

◐ **Trend Break:** Affected variables are not comparable with data from prior years.

2020

- Suspended in-person data collection due to the COVID-19 pandemic. Implemented web data collection (in addition to in-person data collection) once data collection resumed.
- Substance use disorder defined according to DSM-5 criteria rather than DSM-IV criteria.

⊙ **Trend Break:** Data from 2020 onward are not comparable with data from prior years.

2021

- Continued multimode (in-person or web) data collection. Resumed year-round data collection. Revised the person-level analytic weight to account for the two modes.

⊙ **Trend Break:** Data from 2021 onward are not comparable with data from prior years.

2022

- Modernized substance use and mental health treatment questionnaire sections.

◐ **Trend Break:** Substance use and mental health treatment data from 2022 are not comparable with data from prior years.

✖ ✖ Data not publicly available	⊗ ⊗ Data not collected	▣ Data may need adjustments; use caution when comparing across years
⊙ ⊙ Trend break (data not comparable with prior years)	-●●- Data comparable with prior years	◐ Partial trend break (some measures still comparable with prior years)

COVID-19 = coronavirus disease 2019; DSM-IV = *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (American Psychiatric Association, [1994]); DSM-5 = *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition (American Psychiatric Association, [2013]).

[1] See https://www.samhsa.gov/data/data-we-collect/nsduh/datafiles?data_collection=1124&year=2015.

3. NSDUH Population

The target population for NSDUH is the civilian, noninstitutionalized population aged 12 years or older residing within the United States.

The survey includes the following:

- residents of households (e.g., individuals living in houses or townhouses, apartments, and condominiums; civilians living in housing on military bases); and
- individuals in noninstitutional group quarters (e.g., shelters, rooming or boarding houses, college dormitories, migratory workers' camps, halfway houses).

The survey excludes the following:

- individuals with no fixed household address (e.g., homeless and/or transient people not in shelters);
- active-duty military personnel; and
- residents of institutional group quarters, such as jails or hospitals.

Those who are unable to take the survey in either English or Spanish are also excluded.

For a more detailed description of the NSDUH population, see Section 3.2.1, Housing Units, in the 2023 NSDUH Field Interviewer Manual at https://www.samhsa.gov/data/report/nsduh-2023-field-interviewer-manual (CBHSQ, 2022).

4. Sample Design, Data Collection, and Response Rates

4.1 Sample Design

NSDUH utilizes a coordinated, state-based sample design. The sample design includes a multistage area probability sample within each state and the District of Columbia. States were the first level of stratification. Each state was further stratified into approximately equally populated state sampling regions (SSRs). There were then five stages of selection to create the sample:

- Stage 1: Census tracts were selected within each SSR.
- Stage 2: Census block groups were selected within those census tracts.
- Stage 3: Smaller area segments were selected within the census block groups.
- Stage 4: Dwelling units (DUs) were selected within segments to receive a screener.
- Stage 5: Within each selected DU, up to two residents who were at least 12 years old were selected for the interview.

For more information on sample design, see Chapter 2, Coordinated Sample Design for the 2014-2023 NSDUHs, at https://www.samhsa.gov/data/report/nsduh-2023-sample-design-report (CBHSQ, 2024e).

4.2 Data Collection

Beginning with the 2021 NSDUH, multimode data collection was implemented, in which respondents completed the survey in person or via the web. Most NSDUH questions for in-person data collection are self-administered using audio computer-assisted self-interviewing, but topics considered less sensitive are collected by field interviewers. The web version of the survey is entirely self-administered via the internet, and respondents complete the survey using a computer, mobile device, or tablet. The survey is available in English and Spanish.

For more information on data collection, see Section 2.2, Data Collection Methodology and Questionnaire Changes for 2023, at https://www.samhsa.gov/data/report/2023-methodological-summary-and-definitions (CBHSQ, 2024f).

4.3 Response Rates

A final sample of 67,679 interviews was obtained for the 2023 survey. The weighted screening response rate was 24.36 percent, and the weighted interview response rate was 50.45 percent for the 2023 NSDUH.

For more information on demographic-specific response rates, see Section 3.3.1, Screening and Interview Response Rate Patterns, at https://www.samhsa.gov/data/report/2023-methodological-summary-and-definitions (CBHSQ, 2024f).

5. Selection of Variables for Analysis: Variable Naming Conventions and Hierarchy

The NSDUH codebooks are guides to each year's data files. For each variable in a data file, the codebook provides the variable name, a description of the variable, value codes and their meanings, and an unweighted frequency distribution. Many variables originate directly as interview items and include information on questionnaire wording. Other variables are created from more than one variable, in which case the codebook explains how they were constructed.

Generally, variables are named consistently across years when the content of questions is identical or similar enough to be deemed comparable. Only variables that change explicitly are given new names.

5.1 Variable Processing Steps and Types of Variables

Survey data are initially processed to create a raw data file that consists of one record for each interview. Data from this raw file undergo different types of processing to create variables for analysis. The final file used for analysis includes the following types of variables:

- **Edited Variables:** Variables created from the raw data file that are altered after response data are reviewed to identify and address inconsistent data among related variables. Alterations may consist of making logical inferences using other data in the respondent's record, identifying items legitimately skipped, or setting inconsistencies to a missing value.

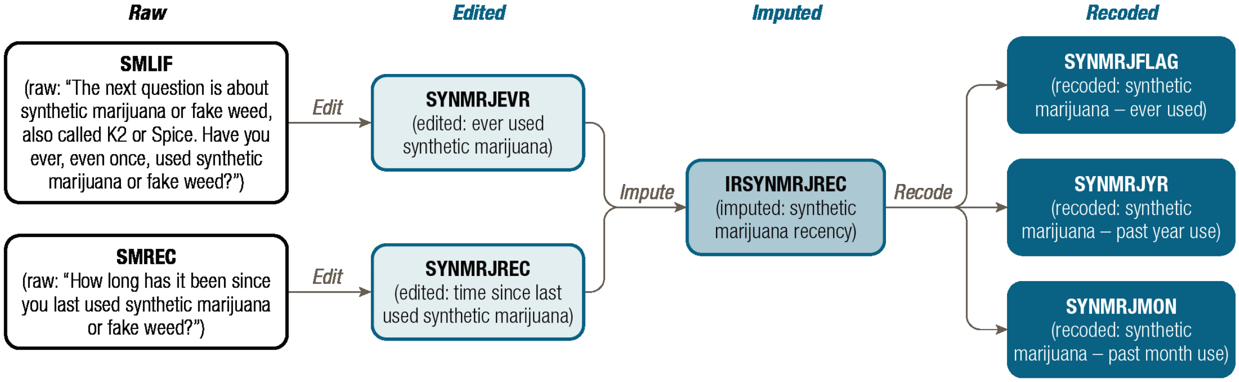
- **Imputed Variables:** Variables that have missing data replaced with nonmissing values using statistical imputation procedures. These variables are identified in the codebook by their labels, which include "IMPUTATION REVISED" or abbreviations such as "IMP REV." Imputed variables are generally found in sections labeled as "Imputed." In addition, most imputed variables have names with the prefix "IR-."
  - Associated indicator variables tell the user which values are imputed and which ones are not. These indicator variables have the words "IMPUTATION INDICATOR" or abbreviations such as "IMP IND" in their labels and are identified with the prefix "II-" in the variable name.
- **Recoded Variables:** Variables created using one or more edited or imputed source variables. These variables are often the versions used in final analysis and can be identified in the codebook by their labels, which have "RC-" preceding the description. Recoded variables are generally found in sections labeled as "Recoded." The missing data codes contained in the source variables are recoded to the standard missing code (.) for recoded variables. It is intended that cases containing these missing codes be excluded from an analysis.
- **Other Variables:** Variables that do not directly relate to the questionnaire and contain supplemental information such as sampling, disclosure, and indicator variables.

For more information on editing and imputation procedures, see Chapters 2 and 3 in the 2022 Editing and Imputation Report at https://www.samhsa.gov/data/report/nsduh-2022-editing-and-imputation-report (CBHSQ, 2024a).

Figure 5.1 shows an example of the lifecycle of the variable processing steps from a raw question to a recoded variable. The raw variable SMLIF corresponds to the questionnaire item "Have you ever, even once, used synthetic marijuana or fake weed?" The raw variable SMREC corresponds to the questionnaire item "How long has it been since you last used synthetic marijuana or fake weed?" These raw variables are directly from the questionnaire; they are not available in the PUF. However, once they go through initial processing, their edited versions are included in the PUF (see edited variables SYNMRJEVR and SYNMRJREC). The two edited variables are evaluated to create one imputed version of the measure (see imputed variable IRSYNMRJREC).[2] This imputed variable is then recoded into analysis variables coded as 1 for "Yes" or 0 for "No" for lifetime, past year, or past month synthetic marijuana use (see recoded variables SYNMRJFLAG, SYNMRJYR, and SYNMRJMON, respectively). Not all edited variables have corresponding imputed or recoded versions. Similarly, imputed variables and recoded variables may derive from one or more edited variables.

[2] All imputation-revised variables are denoted with the prefix "IR-." Recoded variables that are derived from imputation-revised variables and, therefore, do not contain missing values will have the imputed variable(s) listed as source variables.

Figure 5.1 Variable Processing Steps


Raw
Edited
Imputed
Recoded
SMLIF
(raw: "The next question is about synthetic marijuana or fake weed, also called K2 or Spice. Have you ever, even once, used synthetic marijuana or fake weed?")
SMREC
(raw: "How long has it been since you last used synthetic marijuana or fake weed?")
Edit
Edit
SYNMRJEVR
(edited: ever used synthetic marijuana)
SYNMRJREC
(edited: time since last used synthetic marijuana)
Impute
IRSYNMRJREC
(imputed: synthetic marijuana recency)
Recode
SYNMRJFLAG
(recoded: synthetic marijuana – ever used)
SYNMRJYR
(recoded: synthetic marijuana – past year use)
SYNMRJMON
(recoded: synthetic marijuana – past month use)

5.2 Finding and Selection of Analysis Variables

Variables in the codebook are organized in sections by subject matter. Beginning with the demographics section, section names that do not begin with "imputed" or "recoded" contain edited variables and correspond directly to sections of the questionnaire. Where applicable, subsections containing imputation-revised and recoded variables follow their corresponding edited variables. For example, the recoded education section is a subsection under the education section (which contains the edited variables), and the imputed employment section is a subsection under the employment section (which contains the edited variables).

When imputed or recoded variables are provided, users are encouraged to use them to produce estimates rather than using edited variables from the data file. Benefits to using imputed or recoded variables include the following:

- Reducing bias associated with missing data or item nonresponse.
- Reducing the amount of additional coding needed to create an analysis-ready variable.
- Enhancing the integrity and completeness of the dataset, allowing for more accurate and comprehensive statistical analysis.

However, exceptions exist when data users may wish to use edited versions of variables to explore nonresponsiveness or missing data for particular questions. For example, if a user is interested in the percentage of people who did not know whether they used fentanyl among those who used illicit drugs in the past year, the appropriate level corresponding to "don't know" from the edited variable (typically, code 94, 994, or 9994) would be used to create a new variable for analysis.

5.3 Understanding of Variables and the Questionnaire

As much as possible, the codebook shows the key source variables used in creating edited, imputed, and recoded variables. For most edited variables, the codebook indicates in parentheses the question names from the NSDUH instrument that were used to create the edited variables. This allows users to find source questions more easily in the questionnaire and the context and logic that produced the variable.

Codebook Entry for Edited Variable

```
(QD17, QD17A, QD17B)
EDUSCHLGO                Len : 2    NOW GOING TO SCHOOL

                         1 = Yes
                         2 = No
                         11 = Yes (EDUSKPMON = 30)
                         85 = BAD DATA Logically assigned
                         94 = DON'T KNOW
                         97 = REFUSED
                         98 = BLANK (NO ANSWER)
```

Data users may also need to consult the questionnaire to understand the population that is asked the question (also referred to as the "universe") for a given variable. For example, all variables within the Youth Experiences section of the codebook are administered only to respondents aged 12 to 17 years old. Therefore, it would not be possible to conduct analyses focused on adult populations using these variables because there would be no data for respondents aged 18 or older. Information related to the universe can generally be found in the programming specifications presented with the question text in the questionnaire.

Some questions are open-ended and ask respondents to type in a response. For instance, respondents may be asked to write in the names of other hallucinogens that they used in their lifetime. These typed responses are coded into numeric values for analyses. Data taken from these open-ended questions are referred to as "OTHER, Specify" data. Data users should consider fluctuations in the data related to these open-ended questions when analyzing across years. Source variable values for write-in questions or variables with logically assigned values may not exist consistently across years. For example, a recoded variable may document a "Yes" using both source variable values of "1 = Yes" and "3 = Logically Assigned Yes" in the codebook, even if the "3" is not applicable for the current survey year.

6. Use of Analysis Weights and Variance Estimation Variables

6.1 Use of Analysis Weights

The estimates yielded by NSDUH are based on sample survey data rather than on complete data for the entire population, which means the data must be weighted to obtain unbiased estimates for survey outcomes that are representative of the U.S. population. A "final analysis weight" (ANALWT2_C[3]) is calculated for each respondent. The final person-level analysis weight is built from multiple components, which account for probabilities of selection and adjustments for nonresponse, poststratification, and mode of data collection.

For more information on the analysis weights, see Section 2.3.4, Development of Analysis Weights, at https://www.samhsa.gov/data/report/2023-methodological-summary-and-definitions (CBHSQ, 2024f).

For more details on the weight components and sample weighting procedures, see the 2022 Person-Level Sampling Weight Calibration report at https://www.samhsa.gov/data/report/nsduh-2022-person-level-sampling-weight-calibration-report (CBHSQ, 2024b).

6.2 Use of Variance Estimation Variables

For variance estimation, suitable software (e.g., SUDAAN® [RTI International, 2018], SAS® [SAS Institute Inc., 2017], Stata® [StataCorp LP, 2017], R [R Core Team, 2018], and SPSS® [IBM Corp., 2017]) should be used to take the sample design into account. To properly estimate populations and errors, data file users who perform analyses with SUDAAN should first sort the data by the sample design variables VESTR_C[4] (variance estimation [pseudo] stratum) and VEREP (variance estimation [pseudo] primary sampling unit [PSU] within stratum). Commonly used software packages (e.g., SAS, Stata, R, and SPSS) do not require sorting the dataset before inputting into a procedure. These variables then are specified in a software package to automatically account for the sample design when estimating variances and standard errors. See Section 12, Coding Examples, for programming code examples.

When conducting statistical testing, degrees of freedom should be specified. See Section 10.1, Degrees of Freedom, for more information.

7. Statistical Disclosure Avoidance

To protect the confidentiality of respondents, the PUFs are treated using a multistep statistical disclosure limitation method. First, all directly identifying information, geographic identifiers, links between household and respondent, and variables related to interview mode and the period of the year the interview is administered are removed. Second, the data are subsampled to create a subset of the original full analytic data file. After this, the PUF final person-level analysis weights are recalibrated to the full sample totals to reduce bias due to substitution and variance due to subsampling. The sum of the values of ANALWT2_C on the PUF is identical to the sum of the values of the person-level weight on the full analytic file.

Because of these disclosure avoidance steps, estimates derived from the PUFs may not exactly match numbers in reports published by SAMHSA, which were calculated using the RUFs. For more information on statistical disclosure limitation procedures, see Section 3, Confidentiality of Data,

[3] For 2019 and prior survey years, the person-level analysis weight was called ANALWT_C.
[4] For 2019 and prior survey years, the variance estimation variable was called VESTR.

in the PUF Codebook Introduction (CBHSQ, 2024g). Full NSDUH RUFs are available only via the Research Data Center (RDC) at https://www.samhsa.gov/data/data-we-collect/samhsa-rdc.

8. Data Subsetting

Because of its complex sampling design, it is important to subset NSDUH data correctly when working with specific domains (i.e., subpopulations) of interest. Census tracts, which serve as PSUs within sampling strata, would be impacted by incorrectly subsetting NSDUH data. More specifically, if an entire PSU is excluded when subsetting the data, a reduced number of PSUs would be used by the statistical software for variance estimation. Using a reduced number of PSUs will provide either an incorrect variance estimate or an error message. By including all PSUs in a stratum, whether or not the PSU is included in the domain of interest, the statistical analysis software computes the variances appropriately. For more details on the NSDUH variance estimation variables, see Section 6, Use of Analysis Weights and Variance Estimation Variables. Also, when subsetting data to specific domains of interest, data users should be aware of the suppression criteria recommendations for NSDUH (see Section 11, Guidance for Suppressing Unreliable Estimates).

When analyzing NSDUH data among a specific domain, it is generally not recommended to subset to the domain of interest in a data step (i.e., to create a separate dataset including only the domain of interest for further analysis). Rather, it is recommended to load the full sample of data into the software's statistical procedure and to reduce the data to the appropriate subset within the procedure (e.g., SUDAAN SUBPOPN statement [RTI International, 2018], SAS DOMAIN statement, R subset argument in svydesign function, SPSS SELECT IF command within the CSSELECT procedure, Stata subpop() option). This enables the software to recognize the complete data structure for variance estimation, resulting in accurate standard errors for statistical inference.

9. Data Pooling

For certain analyses, sample sizes may not be adequate to support inferences using only a single year of survey data. In these instances, estimates can be produced from annual averages based on combined data from 2 or more survey years. However, due to changes in the NSDUH questionnaire, it is not appropriate to pool data if the variables of interest are not comparable between years. Information related to variable comparability between study years can be found in the appropriate Variable Crosswalk Charts.[5] See Figure 2.1 for details on NSDUH history and trend breaks.

To produce estimates of annual averages of totals based on multiple years of data (e.g., the estimated number of people who used alcohol in the past 3 years), the person-level analysis weight should be divided by the number of years being combined.

[5] See https://www.samhsa.gov/data/data-we-collect/nsduh/datafiles.

Concatenated PUFs with multiple years of NSDUH data may be downloaded for analysis, or users may download individual year PUFs and combine them.[6]

10. Estimation and Statistical Testing

10.1 Degrees of Freedom

The degrees of freedom (*df*) value is needed to determine whether the observed difference between estimates is statistically significant. For the PUF, the *df* are 50 for the 2014 NSDUH and onward national estimates under the current sampling design and 60 for the 2002-2013 NSDUH national estimates under the previous sampling design.

The recommended number of degrees of freedom may depend on the type of analysis. Table 10.1 summarizes the *df* used for key NSDUH analyses per the sample design.

Table 10.1 Key NSDUH Analyses and Degrees of Freedom for the Public Use Data Files per the 2002-2023 NSDUH Sample Design

Analyses	Sample Design Years[1]	Degrees of Freedom for Public Use Data Files
Analyses involving the whole population or a nongeographic subpopulation	2014-2023	50
	2005-2013	60
	2002-2004	60
Analyses for estimates of averages, including mean age at first use	2014-2023	Number of nonempty[2] strata (for each estimate/subpopulation)
	2005-2013	60
	2002-2004	60

[1] The NSDUH sample design variables were revised in 2005 and 2014. The 2005 revisions were applied retroactively to the 1999-2004 surveys. Because of survey improvements in the 2002 NSDUH, the 2002 data constitute a new baseline, so this table does not include information before 2002.

[2] A stratum or primary sampling unit (PSU) is empty for a given subpopulation if the respondent pool contains no subpopulation members in the stratum or PSU. To identify the number of nonempty strata, data users can count the unique levels of the VESTR_C variable within the analysis domain.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2002-2023.

When combining any years of data (e.g., 2022 and 2023), the *df* remain the same as if the set were a single year (e.g., 50 for national estimates) when these years are part of the same sample design. When combining a pair of years with different *df* (e.g., 2013 and 2014), the specific number of *df* can be computed by counting the unique values of VESTR_C. For example, when producing estimates by combining data for 2022 and 2023, *df* = 50 can be used because the design remained the same across those 2 years. When combining data for 2013 and 2014, *df* = 110 because the sample design changed in 2014.

[6] See https://www.samhsa.gov/data/data-we-collect/nsduh/datafiles.

Because of various disclosure control procedures, the *df* differ between the PUFs and RUFs. For more information about how the *df* value is calculated, see Chapter 6, Degrees of Freedom, at https://www.samhsa.gov/data/report/nsduh-2022-statistical-inference-report (CBHSQ, 2024c).

10.2 Statistical Significance of Differences

Statistical tests may be used to compare NSDUH estimates among subdomains (e.g., male vs. female) or across time (e.g., 2022 vs. 2023). Statistical tests evaluate whether observed differences in estimates between groups occur due to random variability or whether they reflect actual differences between the populations being compared.

10.2.1 Comparison of Estimates between Years

Statistical tests can be conducted to compare estimates between years (e.g., past month alcohol use in 2023 vs. 2022). These statistical tests indicate whether an estimate in the first year being compared is lower than, greater than, or similar to the corresponding estimate in the second year being compared. For details on example code for statistical testing between years, see Section 12, Coding Examples, Exhibits 12.2.G, 12.3.G, and 12.4.G.

If comparing outcomes across years, data users should exercise caution when comparing estimates and confirm the ability to compare the measures and years in question. Estimates should never be compared across years when there is a trend break. Variable comparability between study years beginning with the 2002 NSDUH can be found in the appropriate Variable Crosswalk Charts.[7]

Caution is needed when interpreting changes across years in the estimated numbers of people. For more information about considerations for comparing estimates between years, see Section 3.2.3.2, Significance Testing of Estimates in Different Years, at https://www.samhsa.gov/data/report/2023-methodological-summary-and-definitions (CBHSQ, 2024f).

10.2.2 Comparison of Estimates between Categorical Subgroups

Statistical tests can be conducted to compare estimates between population subgroups within a given year (e.g., comparing substance use between adolescents aged 12 to 17 and young adults aged 18 to 25). Comparing population subgroups defined by three or more levels of a categorical variable involves two steps:

1. Log-linear chi-square tests of independence of the subgroup and the prevalence variables should be conducted first to control the error level for multiple comparisons.
2. If Shah's Wald F test (transformed from the standard Wald chi-square) indicates overall significant differences, then the significance of each particular pairwise comparison of interest should be tested using analytic procedures that properly account for the sample design.

[7] See https://www.samhsa.gov/data/data-we-collect/nsduh/datafiles.

This two-step procedure protects against inappropriate inferences being drawn due to the number of pairwise differences tested. For details on example code for statistical testing of population subgroups, see Section 12, Coding Examples, Exhibits 12.1.G, 12.1.H, 12.1.I, 12.2.E, 12.2.F, 12.3.E, 12.3.F, 12.4.E, and 12.4.F.

Significance tests can also be conducted to compare estimates between individual subgroups and the full population (e.g., adults employed full time vs. all adults). Because this testing involves two overlapping domains, a stacked dataset that includes two records for each respondent in the overlap is needed for analysis.

For a coding example on using stacked data for analysis, see Exhibit A.2.10 in the 2022 Statistical Inference Report at https://www.samhsa.gov/data/report/nsduh-2022-statistical-inference-report (CBHSQ, 2024c).

For more information on statistical testing, see Chapter 7, Statistical Significance of Differences, at https://www.samhsa.gov/data/report/nsduh-2022-statistical-inference-report (CBHSQ, 2024c).

10.3 Standard Errors of Totals

To compute the SE of the totals, NSDUH implements two different methods depending on the domain for the estimate:

1. If the domain is forced to match its respective U.S. Census Bureau population estimates through poststratification during the weighting process (Table 10.2), then an alternative SE estimation method is used. The SE of the total = the SE of the controlled domain = the weighted sample size × the SE for the mean/proportion. An example of how this is calculated is included in Section 12, Coding Examples.
2. For all other domains (e.g., those not forced to match the U.S. Census Bureau population estimates), the SE of the total is obtained directly from the statistical analysis software.

Table 10.2 Demographic Domains for the Public Use File That Should Use the Alternative Standard Error (SE) Estimation Method for Calculating Standard Errors of the Estimated Numbers of Individuals (Totals)

Main Effect	Two-Way Interaction[1]
Age Group	
12-17	
18-25	**Age Group × Sex[2]**
26-34	(e.g., males aged 12-17)
35 or Older	
Collapsed Age Group Categories[2]	
Sex	
Male	
Female	

Table 10.2 Demographic Domains for the Public Use File That Should Use the Alternative Standard Error (SE) Estimation Method for Calculating Standard Errors of the Estimated Numbers of Individuals (Totals) (continued)

Main Effect	Two-Way Interaction[1]
Race/Ethnicity	
Hispanic or Latino	
Not Hispanic or Latino, White[3]	

NOTE: The alternative SE estimation method does not affect the SEs for the corresponding means and proportions. These latter SEs are calculated directly in the statistical analysis software, whereas the alternative SE estimation method is computed outside of the statistical analysis software.

NOTE: Domains in this table are specific to the public use files because of different adjustments applied to the weights.

[1] Unless otherwise noted, the domains for the two-way interaction are the same as the main-effect domains (including the collapsed age categories). Two-way interactions involving age group include the main-effect and collapsed age group categories. If age groups are listed in the two-way interaction columns, then only those age groups can be collapsed to form broader age categories.

[2] Main-effect age group categories shown in the table can be collapsed to form broader age group categories (e.g., 12 or older, 18 or older, 18 to 34, 12 to 25). Collapsed main-effect age group categories and the two-way interaction with other main-effect demographic domains shown (e.g., males aged 35 or older) also use the alternative SE estimation method because the collapsed main effects will sum to the census totals for the category being defined. However, broader age groups that include only a subset of the main-effect age groups (e.g., 12 to 20, 21 or older), age groups finer than the main-effect age groups (e.g., 12 and 13, 18 to 20), or two-way interactions of these types of collapsed age categories with other main-effect domains (e.g., females aged 21 or older) should not use the alternative SE estimation method.

[3] This domain is considered a two-way interaction. Thus, any additional domains crossed with Not Hispanic or Latino, White (e.g., Not Hispanic or Latino, White people aged 18 to 25) represent three-way interactions that do not use the alternative SE estimation method.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2023.

10.4 Confidence Intervals

Confidence intervals (CIs) provide a scale to judge how close the sample statistic is likely to be to the true population parameter under repeated sampling. CIs for NSDUH are computed using *df* and critical values of the *t*-distribution, which will change, dependent on which sample of data is being analyzed. In terms of testing for differences between prevalence rates shown with 95 percent CIs, it is important to note that two overlapping 95 percent CIs do not necessarily imply their rates are statistically the same at the 5 percent level of significance.

NSDUH uses the logit transformation method to calculate CIs for proportions. The logit transformation yields asymmetric interval boundaries that are always between 0 and 1 and that are more balanced in terms of whether the true value falls below or above the interval boundaries. This is desirable for NSDUH estimates because many are small percentages that would include values below 0 in the interval without this transformation. For values close to 0, the distribution of a logit-transformed estimate approximates the normal distribution more closely than the standard estimate. For more information on computing CIs, see Chapter 8, Confidence Intervals, at https://www.samhsa.gov/data/report/nsduh-2022-statistical-inference-report (CBHSQ, 2024c). For guidance on computing CIs with various statistical analysis software to correctly analyze NSDUH data, see Section 12, Coding Examples, Exhibits 12.1.D, 12.2.C, 12.3.C, and 12.4.C.

10.5 Regression Analysis

Multiple regression analysis can be implemented for logistic regression if the outcome is a binary variable or for linear regression if the outcome is a continuous variable to study the relationship between an outcome and several covariates. For details on example code for performing regression analysis on NSDUH data, see Section 12, Coding Examples, Exhibits 12.1.K and 12.1.L.

11. Guidance for Suppressing Unreliable Estimates

To reduce the number of estimates with unacceptable levels of statistical reliability, direct estimates from NSDUH may need to be suppressed or omitted from reports or tables. The criteria used to determine suppression of direct estimates in NSDUH data products are based on evaluation of the following components:

- prevalence (for proportion estimates),
- relative standard error (RSE, defined as the ratio of the standard error [SE] over the estimate),
- nominal (actual) sample size, and
- effective sample size for each estimate.

The specific suppression criteria for various NSDUH estimates are summarized in Table 11.1. Data users should suppress estimates if any of these conditions are met.

Table 11.1 Summary of 2023 NSDUH Suppression Rules

Estimate	Suppress if:
Prevalence rate, $\hat{p}$, with nominal sample size, *n*, and design effect, *deff* $\left(deff = \dfrac{n\left[SE(\hat{p})\right]^2}{\hat{p}(1-\hat{p})} \right)$	(1) The estimated prevalence rate, $\hat{p}$, is < .00005 or > .99995, or (2) $\dfrac{\mathrm{SE}(\hat{p})/\hat{p}}{-\ln(\hat{p})} > .175$ when $\hat{p} \le .5$, or $\dfrac{\mathrm{SE}(\hat{p})/(1-\hat{p})}{-\ln(1-\hat{p})} > .175$ when $\hat{p} > .5$, or (3) *Effective n* < 68, where $Effective\ n = \dfrac{n}{deff} = \dfrac{\hat{p}(1-\hat{p})}{\left[SE(\hat{p})\right]^2}$, or (4) *n* < 100. Note: Some unsuppressed estimates of percentages shown in NSDUH data products may appear to be 100.0 because of rounding.
Estimated number (numerator of $\hat{p}$)	The estimated prevalence rate, $\hat{p}$, is suppressed.

Table 11.1 Summary of 2023 NSDUH Suppression Rules (continued)

Estimate	Suppress if:
Means not bounded between 0 and 1 (e.g., mean age at first use), $\overline{x}$, with nominal sample size, *n*	(1) $\mathrm{RSE}(\overline{x}) > .5$, or (2) *n* < 10.

deff = design effect; RSE = relative standard error; SE = standard error.
Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2023.

Data users should be aware that applying these suppression rules to estimates of the same outcome among different domains could result in one estimate being suppressed and another one not being suppressed. In NSDUH data products, this situation would result in the same estimated total being suppressed in one cell but not in another cell. For details on example code for applying the suppression rules to direct estimates, see Section 12, Coding Examples, Exhibits 12.1.M, 12.2.H, 12.3.H, 12.4.H, and 12.5.C.

For more information about suppression criteria, see Section 3.2.2, Suppression Criteria for Unreliable Estimates, at https://www.samhsa.gov/data/report/2023-methodological-summary-and-definitions (CBHSQ, 2024f).

Model-based estimates, such as state or substate small area estimates, may use different criteria to determine suppression. For more information, refer to the year-specific Small Area Estimation Methodology Report at https://www.samhsa.gov/data/nsduh/state-reports.

12. Coding Examples

This section provides examples of SUDAAN, SAS, Stata, R, and SPSS code for generating descriptive statistics, performing statistical tests of independence and tests of differences for means or proportions, conducting domain analyses, calculating CIs, fitting linear or logistic regression models, and analyzing pooled data. These examples are not meant to be exhaustive or to provide instruction for users unfamiliar with a particular software product. Example SUDAAN code is the same for both stand-alone and SAS-callable versions of SUDAAN. Users are advised to apply precision-based suppression rules to estimates as discussed in Section 11, Guidance for Suppressing Unreliable Estimates.

The section number for each programming language, the exhibit number for each example within that section, and a description of the example are provided in Table 12.1.

Table 12.1 Summary of SUDAAN, SAS, Stata, R, and SPSS Exhibits

Section 12.1 SUDAAN	Section 12.2 SAS	Section 12.3 Stata	Section 12.4 R	Section 12.5 SPSS	Description
Exhibit 12.1.A	--	--	--	--	Performs sorting of a dataset.
Exhibit 12.1.B	Exhibit 12.2.A	Exhibit 12.3.A	Exhibit 12.4.A	Exhibit 12.5.A	Creates descriptive statistics for mean estimates.
Exhibit 12.1.C	Exhibit 12.2.B	Exhibit 12.3.B	Exhibit 12.4.B	--	Creates descriptive statistics for proportion estimates.
Exhibit 12.1.D	Exhibit 12.2.C	--	--	--	Calculates confidence intervals using software-specified options.
--	Exhibit 12.2.D	Exhibit 12.3.C	Exhibit 12.4.C	--	Calculates confidence intervals using manual coding.
--	--	Exhibit 12.3.D	Exhibit 12.4.D	Exhibit 12.5.B	Calculates standard errors using manual coding.
Exhibit 12.1.E	--	--	--	--	Performs domain analysis of subset data.
Exhibit 12.1.F	--	--	--	--	Performs analysis of pooled data across years.
Exhibit 12.1.G	Exhibit 12.2.E	Exhibit 12.3.E	Exhibit 12.4.E	--	Performs statistical test of chi-square test of independence.
Exhibit 12.1.H	Exhibit 12.2.F	Exhibit 12.3.F	Exhibit 12.4.F	--	Performs statistical test of pairwise testing.
Exhibit 12.1.I	Exhibit 12.2.G	Exhibit 12.3.G	--	--	Performs statistical test of differences for means.
Exhibit 12.1.J	--	--	Exhibit 12.4.G	--	Performs statistical test of differences for proportions.
Exhibit 12.1.K	--	--	--	--	Performs linear regression modeling.
Exhibit 12.1.L	--	--	--	--	Performs logistic regression modeling.
Exhibit 12.1.M	Exhibit 12.2.H	Exhibit 12.3.H	Exhibit 12.4.H	Exhibit 12.5.C	Creates suppression indicators for estimates by applying the suppression rules.

-- = not available.

12.1 SUDAAN

Before running SUDAAN procedures, the input dataset must be sorted by the nesting variables (VESTR_C and VEREP), or the NOTSORTED option must be used for SUDAAN to create an internal copy of the input dataset properly sorted by the nesting variables. Exhibit 12.1.A displays code that sorts the output dataset by the nesting variables using SAS.

Exhibit 12.1.A Sorting of Data for SUDAAN Analysis

```
PROC SORT DATA=DATANAME;
BY VESTR_C VEREP;
RUN;
```

The SUDAAN procedure DESCRIPT can then be run to produce unweighted sample sizes and weighted (using ANALWT2_C) sample sizes, means and proportions, totals, standard errors (SEs) of means and totals, and *p* values for testing the differences between means, proportions, and totals. Exhibit 12.1.B displays code that computes the weighted sample size (WSUM), unweighted sample size (NSUM), mean (MEAN), weighted total (TOTAL), standard errors of the mean (SEMEAN), and standard errors of the total (SETOTAL) for past month alcohol users (variable ALCMON) in the population.

Because of NSDUH’s complex sample design, estimates are calculated using a method in SUDAAN that is unbiased for linear statistics. This method is based on multistage clustered sample designs where the first-stage (primary) sampling units are drawn with replacement. In SUDAAN, a user must specify DESIGN=WR (meaning with replacement) to invoke the proper Taylor linearized variance estimation method. The option DEFT4 within SUDAAN provides the correct measure of variance inflation due to stratification (or blocking), clustering, and unequal weighting in NSDUH estimation. The following optional additions to the code are displayed in the example:

- The formats defined in the OUTPUT statement increases the length of the calculated statistics.
- The STYLE option shown is preferred for NSDUH national reports to create a more organized output but is optional to the user.
- Design effects for each mean estimate can be output by specifying the DEFFMEAN option in the OUTPUT statement.

Exhibit 12.1.B Descriptive Statistics for Mean Estimates in SUDAAN

```
PROC DESCRIPT DATA=DATANAME DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR_C VEREP;
WEIGHT ANALWT2_C;
VAR ALCMON;
SUBGROUP IRSEX;
LEVELS 2;
TABLES IRSEX;
PRINT WSUM NSUM MEAN SEMEAN TOTAL SETOTAL / REPLACE STYLE=NCHS;
```

Exhibit 12.1.B Descriptive Statistics for Mean Estimates in SUDAAN (continued)

```
OUTPUT WSUM NSUM MEAN SEMEAN TOTAL SETOTAL DEFFMEAN / REPLACE
       NSUMFMT=F8.0 WSUMFMT=F12.0 MEANFMT=F15.10 SEMEANFMT=F15.10
       DEFFMEANFMT=F15.10 TOTALFMT=F12.0 SETOTALFMT=F12.0
       FILENAME="OUT.SUDFILE";
RUN;
```

Whether the SE of the total is taken directly from SUDAAN depends on whether the specified domain (i.e., sex in this example) is fixed as explained in Section 10.3, Standard Errors of Totals. For fixed domains using the dataset produced, an alternative calculation must be used to generate the correct SEs of the total. If using SAS-callable SUDAAN, SAS code can be used to generate the correct estimates. Although two levels were indicated on the LEVELS statement for sex (variable IRSEX), because SUDAAN automatically computes and stores an overall estimate (corresponds to level 0), the alternative calculation performed in SAS is applied to three levels:

```
DATA DATANAME2;
SET OUT.SUDFILE;
IF IRSEX IN (0,1,2) THEN SETOTAL=WSUM*SEMEAN;
RUN;
```

Specifying the SUBGROUP and LEVELS statements stratifies the entire population by sex and produces estimates for each level of the stratification variable, where male = 1 and female = 2. All stratification variables must also be added to the TABLES statement. The following CLASS statement could be used in place of SUBGROUP and LEVELS statements:

```
CLASS IRSEX;
```

Unlike the SUBGROUP and LEVELS statements, multiple CLASS statements can be specified in a single procedure. However, the same variable cannot appear on multiple CLASS statements. A subgroup or stratified analysis using either of these methods differs from a domain analysis, which is shown in Exhibit 12.1.E.

Exhibit 12.1.C displays code that computes estimates for a multilevel categorical variable. Similar to Exhibit 12.1.B, the weighted sample size, unweighted sample size, weighted total, and SE of the total are calculated. However, because the analysis variable has more than two levels, the following modifications must be made:

- The analysis variable is duplicated once for each categorical level (four in this example) on the VAR statement.
- The CATLEVEL option is specified with the analysis variable levels to be shown in the output.
- The options PERCENT and SEPERCENT replace MEAN and SEMEAN in the PRINT and OUTPUT statements.

These changes allow SUDAAN to properly estimate the percentages and SEs for each level of the analysis variable.

Exhibit 12.1.C Descriptive Statistics for Proportion Estimates in SUDAAN

```
PROC DESCRIPT DATA=DATANAME DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR_C VEREP;
WEIGHT ANALWT2_C;
VAR MRJMDAYS MRJMDAYS MRJMDAYS MRJMDAYS;
CATLEVEL 1 2 3 4;
SUBGROUP MRJMON;
LEVELS 1;
TABLES MRJMON;
PRINT WSUM NSUM PERCENT SEPERCENT TOTAL SETOTAL / REPLACE STYLE=NCHS;
OUTPUT WSUM NSUM PERCENT SEPERCENT TOTAL SETOTAL / REPLACE
       FILENAME="OUT.SUDFILE_FREQ";
RUN;
```

To produce CIs, the LOWMEAN and UPMEAN options should be added to the PRINT and OUTPUT statements when estimating means. When estimating CIs for proportions, the options needed on the PRINT and OUTPUT statements are LOWPCT and UPPCT, as shown in Exhibit 12.1.D. By default, SUDAAN generates the confidence limits based on logit transformations.

Exhibit 12.1.D Confidence Interval Options for SUDAAN

```
PRINT MEAN SEMEAN LOWMEAN UPMEAN;
PRINT PERCENT SEPERCENT LOWPCT UPPCT;
```

It is also possible in SUDAAN to use the LOWTOTAL and UPTOTAL options on the PRINT and OUTPUT statements to generate CIs for total estimates. However, as with the calculation of SEs of totals, estimates involving fixed domains require an alternative calculation method to accurately adjust the variances for the domains controlled during the weighting process. PLOWER and PUPPER in the following SAS code are the CIs associated with the mean estimates. These should match what is output by the LOWMEAN and UPMEAN options in SUDAAN. TLOWER AND TUPPER computed in the following code are the CIs associated with the total estimates. These differ from what is provided by the LOWTOTAL and UPTOTAL keywords in SUDAAN.

```
DATA CI;
SET OUT.SUDFILE;
T_QNTILE=TINV(0.975,50);
/*DF of 50 is used in this example based on the 2014-2023 NSDUH
Public Use Files. See Table 10.1 for appropriate DF to use.*/
NUMBER=SEMEAN/(MEAN*(1-MEAN));
L=LOG(MEAN/(1-MEAN));
A=L-T_QNTILE*NUMBER;
B=L+T_QNTILE*NUMBER;

PLOWER=1/(1+EXP(-A));
PUPPER=1/(1+EXP(-B));
```

```
     TLOWER=WSUM*PLOWER;
     TUPPER=WSUM*PUPPER;
     RUN;
```

Domain analyses can be performed using any of the examples in this section by including a SUBPOPN or SUBPOPX statement to subset the population to the domain of interest. Only one of these statements can be used in a single procedure, as shown in Exhibit 12.1.E. Subsetting the dataset in this way, as opposed to creating a smaller dataset outside of the SUDAAN procedure, allows SUDAAN to appropriately compute the variance, as described in Section 8, Data Subsetting.

Exhibit 12.1.E Domain Analysis of Subset Data in SUDAAN

```
PROC DESCRIPT DATA=DATANAME DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR_C VEREP;
WEIGHT ANALWT2_C;
SUBPOPN MRJFLAG=1;
VAR IRMJAGE;
SUBGROUP IRSEX;
LEVELS 2;
TABLES IRSEX;
PRINT MEAN SEMEAN / REPLACE STYLE=NCHS;
RUN;
```

The following code calculates past month alcohol use estimates by sex and year using pooled data. Data users need to create an appropriate pooled weight, as described in Section 9, Data Pooling. Exhibit 12.1.F displays code that shows the creation of an appropriate weight using SAS for data pooled across 2 years, followed by the descriptive analysis in SUDAAN. The estimates calculated represent annual averages of combined years of data.

Exhibit 12.1.F Analysis of Pooled Data across Years in SUDAAN

```
DATA DATANAME;
SET TWOYEAR_DATA;
POOLED_WEIGHT = ANALWT2_C/2;
RUN;

PROC DESCRIPT DATA=DATANAME DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR_C VEREP;
WEIGHT POOLED_WEIGHT;
VAR ALCMON;
SUBGROUP YEAR IRSEX;
LEVELS 2 2;
TABLES YEAR*IRSEX;
PRINT WSUM NSUM MEAN SEMEAN TOTAL SETOTAL / REPLACE STYLE=NCHS;
```

Exhibit 12.1.F Analysis of Pooled Data across Years in SUDAAN (continued)

```
OUTPUT WSUM NSUM MEAN SEMEAN TOTAL SETOTAL DEFFMEAN /REPLACE
       NSUMFMT=F8.0 WSUMFMT=F12.0 MEANFMT=F15.10 SEMEANFMT=F15.10
       DEFFMEANFMT=F15.10 TOTALFMT=F12.0 SETOTALFMT=F12.0
       FILENAME="OUT.SUDFILE";

RUN;
```

Exhibit 12.1.G displays code that shows how to create a log-linear chi-square test of independence for a subpopulation defined by three or more levels of a categorical variable. This code calculates Shah's Wald *F* test based on a log-linear model to indicate whether a statistically significant association exists between two variables (i.e., employment status and past month cigarette use in this example). Performing statistical tests in SUDAAN requires the specification of degrees of freedom using the DDF option.

Exhibit 12.1.G Chi-Square Test of Independence in SUDAAN

```
PROC CROSSTAB DATA=DATANAME DDF=50 DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR_C VEREP;
WEIGHT ANALWT2_C;
CLASS CIGMON;
SUBGROUP IRWRKSTAT18;
LEVELS 4;
TABLES IRWRKSTAT18*CIGMON;
TEST LLCHISQ / WALDF;
SETENV DECWIDTH=4 COLWIDTH=15;
PRINT NSUM WSUM TOTPER ROWPER COLPER STESTVAL SPVAL SDF / REPLACE
       STYLE=NCHS;
OUTPUT STESTVAL SPVAL SDF / REPLACE FILENAME="TEST_CHI";
RUN;
```

Once an association between two variables is determined to be significant, pairwise testing can be done to determine individual significance tests across levels of an independent variable. Exhibit 12.1.H displays code that shows significance is tested across levels of employment status using the IRWRKSTAT18 variable.

Exhibit 12.1.H Pairwise Testing in SUDAAN

```
PROC DESCRIPT DATA=DATANAME DDF=50 DESIGN=WR FILETYPE=SAS;
NEST VESTR_C VEREP;
WEIGHT ANALWT2_C;
VAR CIGMON;
SUBGROUP IRWRKSTAT18;
LEVELS 4;
PAIRWISE IRWRKSTAT18 / NAME="Tests of differences for all levels";
PRINT WSUM NSUM MEAN SEMEAN TOTAL SETOTAL T_MEAN P_MEAN / REPLACE
       STYLE=NCHS;
```

Exhibit 12.1.H Pairwise Testing in SUDAAN (continued)

```
OUTPUT WSUM NSUM MEAN SEMEAN TOTAL SETOTAL T_MEAN P_MEAN / REPLACE
       NSUMFMT=F8.0 WSUMFMT=F12.0 MEANFMT=F15.10 SEMEANFMT=F15.10
       TOTALFMT=F12.0 SETOTALFMT=F12.0 FILENAME="OUT.SUDTESTS";
RUN;
```

Exhibit 12.1.I displays code that performs statistical tests of differences between levels of the SUBGROUP variable IRSEX. The T_MEAN and P_MEAN options output the test statistic and the two-sided *p* value from a *t* distribution, respectively, for the test of differences in means of past month alcohol use among males versus females, as specified in the DIFFVAR statement. For a coding example on test of differences for totals, see Exhibit A.2.6 in the 2022 Statistical Inference Report at https://www.samhsa.gov/data/report/nsduh-2022-statistical-inference-report (CBHSQ, 2024c).

Exhibit 12.1.I Tests of Differences for Means in SUDAAN

```
PROC DESCRIPT DATA=DATANAME DDF=50 DESIGN=WR FILETYPE=SAS;
NEST VESTR_C VEREP;
WEIGHT ANALWT2_C;
VAR ALCMON;
SUBGROUP IRSEX;
LEVELS 2;
DIFFVAR IRSEX=(1 2) / NAME="MALES vs FEMALES";
PRINT WSUM NSUM MEAN SEMEAN T_MEAN P_MEAN / REPLACE STYLE=NCHS;
OUTPUT WSUM NSUM MEAN SEMEAN T_MEAN P_MEAN / REPLACE NSUMFMT=F8.0
       WSUMFMT=F12.0 MEANFMT=F15.10 SEMEANFMT=F15.10
       FILENAME="OUT.SUDTESTS";
RUN;
```

Alternatively, the same test can be performed using a CONTRAST statement (in place of the DIFFVAR statement). However, instead of using the variable levels, users must specify a more general linear contrast:

```
CONTRAST IRSEX=(1 -1) / NAME="MALES vs FEMALES";
```

For a multilevel categorical analysis variable, different options are needed to produce the test statistic and *p* value for tests of differences between percents. Exhibit 12.1.J displays code that shows the T_MEAN and P_MEAN options are replaced with T_PCT and P_PCT.

Exhibit 12.1.J Tests of Differences for Proportions in SUDAAN

```
PROC DESCRIPT DATA=DATANAME DDF=50 DESIGN=WR FILETYPE=SAS;
NEST VESTR_C VEREP;
WEIGHT ANALWT2_C;
VAR MRJMDAYS MRJMDAYS MRJMDAYS MRJMDAYS;
CATLEVEL 1 2 3 4;
SUBGROUP IRSEX;
LEVELS 2;
```

Exhibit 12.1.J Tests of Differences for Proportions in SUDAAN (continued)

```
DIFFVAR IRSEX=(1 2) / NAME="MALES vs FEMALES";
PRINT WSUM NSUM PERCENT SEPERCENT T_PCT P_PCT / REPLACE STYLE=NCHS;
OUTPUT WSUM NSUM PERCENT SEPERCENT T_PCT P_PCT / REPLACE
       FILENAME="OUT.SUDFILE_FREQ";
RUN;
```

To perform a test of differences between comparable years of NSDUH data, analysts would need to add the YEAR variable to the SUBGROUP and LEVELS statements. The following code computes the difference between the mean or percentage estimates for YEAR 1 and YEAR 4 for each level of the IRSEX variable and overall. For testing of multiple pairs of years, several DIFFVAR statements could be used.

```
SUBGROUP YEAR IRSEX;
LEVELS 4 2;
DIFFVAR YEAR=(1 4) / NAME="YEAR 1 vs YEAR 4";
```

Exhibit 12.1.K displays code that shows a linear regression model using the continuous variable BMI2 as the outcome variable and the categorical variables CATAGMH2, IRSEX, NEWRACE2, and INCOME as the predictors. The reference levels are specified for three out of the four predictor variables using the REFLEVEL statement. The code computes the beta estimates (BETA), standard errors of the betas (SEBETA), test statistic (T_BETA), and *p* value (P_BETA) from a *t* distribution.

Exhibit 12.1.K Linear Regression in SUDAAN

```
PROC REGRESS DATA=DATANAME DDF=50 DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR_C VEREP;
WEIGHT ANALWT2_C;
REFLEVEL CATAGMH2=1 NEWRACE2=1 INCOME=1;
SUBGROUP CATAGMH2 IRSEX NEWRACE2 INCOME;
LEVELS 4 2 7 4;
MODEL BMI2=CATAGMH2 IRSEX NEWRACE2 INCOME;
SETENV DECWIDTH=6 COLWIDTH=18;
PRINT BETA="BETA" SEBETA="STDERR" DEFT="DESIGN EFFECT"
       T_BETA="T:BETA=0" P_BETA="P-VALUE" / TESTS=DEFAULT
       T_BETAFMT=F8.2 WALDCHIFMT=f6.2 DFFMT=f7.0;
OUTPUT BETA SEBETA T_BETA P_BETA / REPLACE
       FILENAME="OUT.MODEL_OUTPUT";

RUN;
```

Exhibit 12.1.L displays code that fits a logistic regression model using the dichotomous variable ALCMON (with values of 0 and 1) as the outcome variable and the categorical variables CATAGMH2, IRSEX, NEWRACE2, and INCOME as the predictors. The option specifications for logistic regression are the same as the linear regression model. For logistic regression models, the odds ratio and relative risk are often of particular interest. The RISK=ALL option in the PRINT statement generates these estimates.

Exhibit 12.1.L Logistic Regression in SUDAAN

```
PROC RLOGIST DATA=DATANAME DDF=50 DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR_C VEREP;
WEIGHT ANALWT2_C;
REFLEVEL CATAGMH2=1 NEWRACE2=1 INCOME=1;
SUBGROUP CATAGMH2 IRSEX NEWRACE2 INCOME;
LEVELS 4 2 7 4;
MODEL ALCMON=CATAGMH2 IRSEX NEWRACE2 INCOME;
SETENV DECWIDTH=6 COLWIDTH=18;
PRINT BETA="BETA" SEBETA="STDERR" DEFT="DESIGN EFFECT"
        T_BETA="T:BETA=0" P_BETA="P-VALUE" / RISK=ALL
        TESTS=DEFAULT T_BETAFMT=F8.2 WALDCHIFMT=f6.2 ORFMT=f10.2
        LOWORFMT=f10.2 UPORFMT=f10.2 DFFMT=f7.0;
OUTPUT BETA SEBETA T_BETA P_BETA / REPLACE
        FILENAME="OUT.MODEL_OUTPUT";

RUN;
```

As discussed in Section 11, Guidance for Suppressing Unreliable Estimates, estimates with unacceptable levels of statistical reliability should not be reported. Exhibit 12.1.M displays code that applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1 using output from SUDAAN Exhibit 12.1.B. Manual coding of calculations is required to create the various conditions of the suppression rule in SAS. The necessary steps are included as commented text in the example code.

Exhibit 12.1.M Manual Application of Suppression Rule in SAS Using SUDAAN Output

```
DATA ESTIMATE;
SET OUT.SUDFILE;
/******APPLY THE PREVALENCE ESTIMATE SUPPRESSION RULE*******/
/* CALCULATE THE RELATIVE STANDARD ERROR */
IF MEAN GT 0.0 THEN RSE=SEMEAN/MEAN;

/* CALCULATE THE RELATIVE STANDARD ERROR OF NATURAL LOG P */
IF 0.0 LT MEAN LE 0.5 THEN RSELNP=RSE/ABS(LOG(MEAN));
ELSE IF 0.5 LT MEAN LT 1.0 THEN
RSELNP=RSE*(MEAN/(1-MEAN))/(ABS(LOG(1-MEAN)));

/*CALCULATE THE EFFECTIVE SAMPLE SIZE*/
EFFNSUM=NSUM/DEFFMEAN;
```

Exhibit 12.1.M Manual Application of Suppression Rule in SAS Using SUDAAN Output (continued)

```
/*SUPPRESSION RULE FOR PREVALENCE ESTIMATES*/
IF (MEAN LT 0.00005) OR (MEAN GT 0.99995) OR (RSELNP GT 0.175) OR
        (EFFNSUM < 68)
OR (NSUM <100) THEN SUPRULE=1;

/*SUPPRESSION RULE FOR MEANS NOT BOUNDED BY 0 AND 1, I.E. AVERAGES
        (COMMENTED OUT FOR THIS EXAMPLE)*/
/*IF (RSE GT 0.5) OR (NSUM < 10) THEN SUPRULE=1;*/
RUN;
```

12.2 SAS

SAS SURVEY procedures do not require either data sorting or specifying the degrees of freedom before analysis of complex survey data. In SAS, data users must use two separate statements (STRATA and CLUSTER) within an appropriate procedure to properly incorporate the nesting variables for the survey design. Exhibit 12.2.A displays code that computes the prevalence estimates, unweighted and weighted sample sizes, weighted and unweighted total, and SEs for the weighted estimates of past month alcohol users (variable ALCMON) by sex in the population. The default variance estimation is the Taylor series linearization method.

Exhibit 12.2.A Descriptive Statistics for Mean Estimates in SAS

```
PROC SURVEYMEANS DATA=DATANAME SUMWGT NOBS MEAN SUM;
STRATA VESTR_C;
CLUSTER VEREP;
WEIGHT ANALWT2_C;
VAR ALCMON;
DOMAIN IRSEX;
ODS OUTPUT DOMAIN=OUT.SASFILE;
RUN;
```

Whether the SE of the total is taken directly from SAS depends on whether the specified domain (i.e., sex in this example) is fixed, as explained in Section 10.3, Standard Errors of Totals. For output generated by the SAS PROC SURVEYMEANS procedure, an overall estimate is not stored in the same dataset as the estimates for levels of IRSEX, as it was in SUDAAN. Therefore, the alternative calculation is applied only to the two levels of the IRSEX variable:

```
DATA DATANAME2;
SET OUT.SASFILE;
IF IRSEX IN (1,2) THEN SETOTAL=SUMWGT*STDERR;
RUN;
```

Exhibit 12.2.B displays code that shows how to compute the mean, SE of the mean, weighted and unweighted sample size, weighted total, and SE of the total corresponding to levels of a categorical variable.

Exhibit 12.2.B Descriptive Statistics for Proportion Estimates in SAS

```
PROC SURVEYFREQ DATA=DATANAME;
WHERE MRJMON=1;
STRATA VESTR_C;
CLUSTER VEREP;
WEIGHT ANALWT2_C;
TABLES MRJMDAYS;
RUN;
```

For SAS, Wald confidence limits are the default for both PROC SURVEYMEANS and PROC SURVEYFREQ. To produce the logit confidence limits, the CL option must be specified with the TYPE=LOGIT modifier in the TABLES statement of PROC SURVEYFREQ. In PROC SURVEYMEANS, the CLM option can be specified to generate the confidence interval of the mean, as shown in Exhibit 12.2.C. However, there is no option to change the calculation method from Wald to logit confidence limits. A logit transformation would have to be manually coded.

Exhibit 12.2.C Confidence Interval Options for SAS

```
PROC SURVEYMEANS DATA=DATANAME SUMWGT NOBS MEAN SUM CLM;
TABLES MRJMDAYS / CL (TYPE=LOGIT);
```

SAS does not provide options to generate CIs for the totals, regardless of whether the domain is fixed. Exhibit 12.2.D displays code that uses output from Exhibit 12.2.A to create the 95 percent CIs for means and totals using manual coding steps for calculations in SAS. The confidence interval associated with the total estimate generated by this code is for fixed domains. The variables PLOWER and PUPPER are the CIs associated with the mean. The variables TLOWER and TUPPER are the CIs associated with the weighted total.

Exhibit 12.2.D Manual Coding for Calculation of Confidence Interval in SAS

```
DATA CI;
SET OUT.SASFILE;
T_QNTILE=TINV(0.975,50);
/*DF of 50 is used in this example based on for the 2014-2023 NSDUH
       Public Use Files. See Table 10.1 for appropriate DF to
       use.*/
NUMBER=STDERR/(MEAN*(1-MEAN));
L=LOG(MEAN/(1-MEAN));
A=L-T_QNTILE*NUMBER;
B=L+T_QNTILE*NUMBER;

PLOWER=1/(1+EXP(-A));
PUPPER=1/(1+EXP(-B));

TLOWER=SUMWGT*PLOWER;
TUPPER=SUMWGT*PUPPER;
RUN;
```

Exhibit 12.2.E displays code that shows how to create a log-linear chi-square test of independence for a subpopulation defined by a multilevel categorical variable. This code calculates both a Rao-Scott chi-square test and the log-linear chi-square Wald *F* test to indicate whether past month cigarette use (variable CIGMON) is associated with employment status.

Exhibit 12.2.E Chi-Square Test of Independence in SAS

```
PROC SURVEYFREQ DATA=DATANAME;
WHERE IRWRKSTAT18 IN (1,2,3,4);
STRATA VESTR_C;
CLUSTER VEREP;
WEIGHT ANALWT2_C;
TABLE IRWRKSTAT18*CIGMON / COL ROW CHISQ WLLCHISQ;
ODS OUTPUT WLLCHISQ=OUT.SAS_CHI;
RUN;
```

Once an association is determined by a significance chi-square test, then pairwise testing, as shown in Exhibit 12.2.F, can be done to determine individual significance tests across levels of employment status.

Exhibit 12.2.F Pairwise Testing in SAS

```
PROC SURVEYREG DATA=DATANAME;
WHERE IRWRKSTAT18 IN (1,2,3,4);
STRATA VESTR_C;
CLUSTER VEREP;
WEIGHT ANALWT2_C;
CLASS IRWRKSTAT18;
MODEL CIGMON=IRWRKSTAT18 / NOINT VADJUST=NONE SOLUTION;
LSMEANS IRWRKSTAT18 / DIFF;
RUN;
```

Exhibit 12.2.G displays code that calculates tests of differences for a dichotomous variable across 2 comparable years. The option NOINT omits the intercept from the model and the option VADJUST=NONE specifies that no variance adjustment is used. The SOLUTION option displays the betas or parameter estimates.

Exhibit 12.2.G Tests of Differences for Means in SAS

```
PROC SURVEYREG DATA=TWOYEAR_DATA;
STRATA VESTR_C;
CLUSTER VEREP;
WEIGHT ANALWT2_C;
DOMAIN IRSEX;
CLASS YEAR;
MODEL ALCMON = YEAR / NOINT VADJUST=NONE SOLUTION;
LSMEANS YEAR / DIFF;
ODS OUTPUT DIFFS = OUT.SASTESTS;
RUN;
```

Exhibit 12.2.H displays code that applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1. Because SAS does not automatically generate the SE or design effect, both of which are necessary in the manual calculation, additional data manipulation steps are necessary to apply the suppression rule using SAS data. The code uses the dataset output from Exhibit 12.2.A and merges it with the results of the PROC UNIVARIATE step to apply the rules. Instructions for each step are provided in the commented text.

Exhibit 12.2.H Manual Application of Suppression Rule in SAS

```
/*Sort analysis dataset by domain variables*/
PROC SORT DATA=DATANAME; BY IRSEX; RUN;

/*Calculate the variance under simple random sampling*/
PROC UNIVARIATE DATA=DATANAME VARDEF=WGT;
VAR ALCMON;
WEIGHT ANALWT2_C;
BY IRSEX;
ODS OUTPUT MOMENTS=SASUNI;
RUN;

/*Manipulate dataset output from PROC UNIVARIATE to keep only the
       domain variables and the standard error*/
DATA DEFF (RENAME = (NVALUE1 = SESRS) KEEP = IRSEX NVALUE1);
SET SASUNI;
WHERE LABEL1 = "Std Deviation";
RUN;

/*Sort the output dataset from Exhibit 12.2.A*/
PROC SORT DATA = OUT.SASFILE; BY IRSEX; RUN;

/*Merge the DEFF dataset with the output dataset from Exhibit
       12.2.A*/
DATA SASEST_MERGE;
MERGE OUT.SASFILE DEFF;
BY IRSEX;
RUN;

/*Calculate DEFF of the mean*/
DATA SASEST;
SET SASEST_MERGE;
DEFFMEAN = (STDERR/SESRS)**2*(N-1);

/******APPLY THE PREVALENCE ESTIMATE SUPPRESSION RULE*******/
/* CALCULATE THE RELATIVE STANDARD ERROR */
IF MEAN GT 0.0 THEN RSE=STDERR/MEAN;
```

Exhibit 12.2.H Manual Application of Suppression Rule in SAS (continued)

```
/* CALCULATE THE RELATIVE STANDARD ERROR OF NATURAL LOG P */
IF 0.0 LT MEAN LE 0.5 THEN RSELNP=RSE/ABS(LOG(MEAN));
ELSE IF 0.5 LT MEAN LT 1.0 THEN RSELNP=RSE*(MEAN/(1-
MEAN))/(ABS(LOG(1-MEAN)));

/*CALCULATE THE EFFECTIVE SAMPLE SIZE*/
EFFNSUM=N/DEFFMEAN;

/*SUPPRESSION RULE FOR PREVALENCE ESTIMATES*/
IF (MEAN LT 0.00005) OR (MEAN GT 0.99995) OR (RSELNP GT 0.175) OR
        (EFFNSUM < 68) OR
(N <100) THEN SUPRULE=1;

/*SUPPRESSION RULE FOR MEANS NOT BOUNDED BY 0 AND 1, I.E. AVERAGES
        (COMMENTED OUT FOR THIS EXAMPLE)*/
/*IF (RSE GT 0.5) OR (N < 10) THEN SUPRULE=1;*/
RUN;
```

12.3 Stata

Stata does not require sorting by the nesting variables before proceeding with the analysis.

Exhibit 12.3.A displays code that demonstrates how to compute various types of estimates for past month alcohol use by sex using the Stata svy: mean and svy: total commands. Code is included to compute the prevalence estimate (i.e., mean), SE of the mean, weighted sample size, unweighted sample size, weighted total, and SE of the total. Whether the SE of the total is taken directly from Stata depends on whether the specified domain (i.e., sex in this example) is fixed, as described in Section 10.3, Standard Errors of Totals.

Exhibit 12.3.A Descriptive Statistics for Mean Estimates in Stata

```
use using ".\\dataname.dta", clear
rename *, lower
svyset verep [pw=analwt2_c], strata(vestr_c) dof(50)

gen total_out=.
Gen setotal=.
Gen mean_out=.
Gen semean=.
Gen nsum=.
Gen wsum=.
Gen deffmean=.
```

Exhibit 12.3.A Descriptive Statistics for Mean Estimates in Stata (continued)

```
svy: mean alcmon, over(irsex)
matrix M=e(b)
matrix S=e(V)
matrix N=e(_N)
matrix W=e(_N_subp)

estat effects, deff srssubpop
matrix D=e(deff)

forvalues j=1/2 {
       replace mean_out=(M[1,`j’]) if irsex==`j’
       replace semean=(sqrt(S[`j’,`j’])) if irsex==`j’
       replace nsum=(N[1,`j’]) if irsex==`j’
       replace wsum=(W[1,`j’]) if irsex==`j’
       replace deffmean=(D[1,`j’]) if irsex==`j’
   }

svy: total alcmon, over(irsex)

matrix M=e(b)
matrix S=e(V)

forvalues j=1/2 {
    replace total_out=(M[1,`j’]) if irsex==`j’
    replace setotal=(sqrt(S[`j’,`j’])) if irsex==`j’
   }

keep wsum mean_out semean total_out setotal nsum deffmean irsex

duplicates drop irsex, force

format wsum %-12.0fc
format mean_out %-15.10f
format semean %-15.10f
format total_out %-12.0fc
format setotal %-12.0fc
format nsum %-8.0fc
format deffmean %-15.10f

list irsex wsum nsum mean_out semean total_out setotal
```

Exhibit 12.3.B displays code that shows how to compute estimates corresponding to levels of a categorical variable.

Exhibit 12.3.B Descriptive Statistics for Proportion Estimates in Stata

```
use using ".\\dataname.dta", clear
rename *, lower
svyset verep [pw=analwt2_c], strata(vestr_c) dof(50)
svy: proportion mrjmdays, subpop(mrjmon)
```

Exhibit 12.3.C displays code that computes a 95 percent confidence interval by using the output data from Stata.

Exhibit 12.3.C Manual Coding for Calculation of Confidence Interval in Stata Using Stata Output

```
generate t_qntile = invt(50,0.975)
/*DF of 50 is used in this example based on for the 2014-2023 NSDUH
       Public Use Files. See Table 10.1 for appropriate DF to use.*/
generate number = semean/(mean_out*(1-mean_out))
generate l=log(mean_out/(1-mean_out))
generate a = l-t_qntile*number
generate b = l+t_qntile*number
generate plower = 1/(1+exp(-a))
generate pupper = 1/(1+exp(-b))
generate tlower = wsum*plower
generate tupper = wsum*pupper
duplicates drop year irsex, force

keep year irsex nsum wsum mean_out semean total_out setotal
///t_qntile number l a b plower pupper tlower tupper
```

Exhibit 12.3.D displays code that computes the SE of the total for fixed domains. Because sex is a fixed domain, the SE of the totals would not be taken directly from the output of the previous example but rather would be computed using the alternative SE estimation method.

Exhibit 12.3.D Manual Coding for the Calculation of Standard Errors in Stata Using Stata Output

```
generate setotal2=wsum*semean
replace setotal = setotal2 if inlist(irsex,1,2)
```

Exhibit 12.3.E displays code that shows how to create a log-linear chi-square test of independence for a subpopulation defined by three or more levels of a categorical variable. This code calculates Shah’s Wald *F* test to indicate whether cigarette use is associated with employment status. The output provides both the adjusted and nonadjusted Wald *F* test statistic.

Exhibit 12.3.E Chi-Square Test of Independence in Stata

```
use using ".\\dataname.dta", clear
rename *, lower
generate subpop = 1 if inlist(irwrkstat18,1,2,3,4)
svyset verep [pw=analwt2_c], strata(vestr_c) dof(50)
svy, subpop(subpop): tab cigmon irwrkstat18, llwald noadjust
```

Once an association is determined by a significance Wald *F* test, then pairwise testing, as shown in Exhibit 12.3.F, can be done to determine individual significance tests across levels of employment status.

Exhibit 12.3.F Pairwise Testing in Stata

```
use using ".\\dataname.dta", clear
rename *, lower
generate subpop = 1 if inlist(irwrkstat18,1,2,3,4)
svyset verep [pw=analwt2_c], strata(vestr_c) dof(50)
svy: mean cigmon, over(irwrkstat18)
matrix Me = e(b)
local max=4
matrix output = J(6,7,.)
local counter1 = `max’ - 1
local counter2 = `max’ - 1
local contrast = 0
forvalues i=1/`counter1’ {
local stop = `max’ - `i’ + 1
forvalues j=1/`counter2’ {
local contrast = `contrast’ + 1
test [cigmon]`j’ = [cigmon]`stop’, nosvyadjust ///
matvlc(mtest`contrast’)
matrix output[`contrast’, 1] = `j’
matrix output[`contrast’, 2] = `stop’
matrix output[`contrast’,7]=r(p)
matrix output[`contrast’,4]=sqrt((mtest`contrast’[1,1]))
matrix output[`contrast’,3]=Me[1,`j’]-Me[1,`stop’]
}
local counter2 = `counter2’ - 1
}
svy: total cigmon, over(irwrkstat18)
matrix M = e(b)
```

Exhibit 12.3.F Pairwise Testing in Stata (continued)

```
local max=4
local counter1 = `max’ - 1
local counter2 = `max’ - 1
local contrast = 0
forvalues i=1/`counter1’ {
local stop = `max’ - `i’ + 1
forvalues j=1/`counter2’ {
local contrast = `contrast’ + 1
           test [cigmon]`j’ = [cigmon]`stop’, nosvyadjust ///
           matvlc(test`contrast’)
           matrix output[`contrast’,6]=sqrt((test`contrast’[1,1]))
           matrix output[`contrast’,5]=M[1,`j’]-M[1,`stop’]
      }
local counter2 = `counter2’ - 1
}
matrix colnames output = level1 level2 mean semean total_out ///
setotal mean_pval
matrix list output
```

Exhibit 12.3.G displays code that performs significance testing between comparable years of NSDUH data using output from SUDAAN. The input dataset would need to include at least 2 years of NSDUH data, but Stata does not require the data to be sorted by nesting variables.

Exhibit 12.3.G Tests of Differences for Means in Stata

```
use using ".\\dataname.dta", clear
rename *, lower
svyset verep [pweight=analwt2_c], strata(vestr_c) dof(50)
{
svy: mean alcmon, over(year irsex)
local max=2*2
local range=2
local compmin=`max’-`range’
gen pmean=.
local counter=1
forvalues i=1/1 {
        local counter2=1
        forvalues j=1/2 {
                local stop=`counter2’+`compmin’
                test [alcmon]_subpop_`counter’ = ///
                [alcmon]_subpop_`stop’, nosvyadjust
                replace pmean=r(p) if year==`i’ & irsex==`j’
                local counter=`counter’+1
                local counter2=`counter2’+1
                }
        }
}
```

Exhibit 12.3.G Tests of Differences for Means in Stata (continued)

```
svy: total alcmon, over(year irsex)
{
matrix M = e(b)
local max=2*2
local range=2
local compmin=`max’-`range’
gen total=.
gen setotal=.
local counter=1
forvalues i=1/1 {
        local counter2=1
        forvalues j=1/2 {
                local stop=`counter2’+`compmin’
                test [alcmon]_subpop_`counter’ = ///
                [alcmon]_subpop_`stop’, nosvyadjust
                     matvlc(test`counter’)
                replace setotal= sqrt((test`counter’[1,1])) ///
                if year==`i’ & irsex==`j’
                replace total=M[1,`counter’]-M[1,`stop’] ///
                if year==`i’ & irsex==`j’
                local counter=`counter’+1
                local counter2=`counter2’+1
                }
        }
}

keep irsex total setotal pmean
duplicates drop irsex total setotal pmean, force
drop if total ==.

format pmean %-15.10f
format total %-12.0fc
format setotal %-12.0fc

list irsex total setotal pmean
```

For testing of multiple years versus the current year, more years could be included in the dataset, and the number of tests conducted can be increased by changing the number of “for loops.” For example, conducting tests of differences in means for each year versus the current year for means requires the following code changes to be made:

```
local max=x*2
local range=2
local compmin=`max’-`range’
gen pmean=.
local counter=1
forvalues i=1/(x-1) {
        local counter2=1
        forvalues j=1/2 {
               local stop=`counter2’+`compmin’
               test [alcmon]_subpop_`counter’ = ///
               [alcmon]_subpop_`stop’, nosvyadjust
               replace pmean=r(p) if year==`i’ & irsex==`j’
               local counter=`counter’+1
               local counter2=`counter2’+1
               }
                      }
               }
```

Exhibit 12.3.H displays code that applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1 (commented out in the example) using the data produced by Stata.

Exhibit 12.3.H Manual Application of Suppression Rule in Stata

```
/******APPLY THE PREVALENCE ESTIMATE SUPPRESSION RULE*******/
/*CALCULATE THE RELATIVE STANDARD ERROR*/
generate rse=.
replace rse=semean/mean_out ///
if mean_out > 0.0 & !missing(mean_out)

/* CALCULATE THE RELATIVE STANDARD ERROR OF NATURAL LOG P */
generate rselnp=.
replace rselnp=rse/(abs(log(mean_out))) ///
if mean_out <= 0.5 & mean_out > 0.0
replace rselnp=rse*(mean_out/(1-mean_out)) ///
/(abs(log(1-mean_out))) if mean_out < 1.0 & mean_out > 0.5

/*CALCULATE THE EFFECTIVE SAMPLE SIZE*/
generate effnsum=nsum/deffmean

/*SUPPRESSION RULE FOR PREVALENCE ESTIMATES*/
generate suprule1a=1 if rselnp > 0.175 & !missing(rselnp)
generate suprule1b=1 if mean_out <.00005 & !missing(mean_out)
generate suprule1c=1 if mean_out >.99995 & !missing(mean_out)
generate suprule2=1 if effnsum < 68 & !missing(nsum)
generate suprule3=1 if nsum < 100 & !missing(nsum)
generate suppress=0
replace suppress=1 if suprule1a==1 | suprule1b==1 | ///
suprule1c==1 | suprule2==1 | suprule3==1

/*SUPPRESSION RULE FOR MEANS NOT BOUNDED BY 0 AND 1, I.E. AVERAGES
        (COMMENTED OUT FOR THIS EXAMPLE)*/
/*generate suprule=1 if (nsum < 10 & !missing(nsum))///
| (rse > 0.5 & !missing(rse))*/
```

12.4 R

The R code examples in this section provide guidance on how to use various programming options to produce estimates for NSDUH using various statistical procedures. Note that R code is case sensitive. The examples shown use the following R packages: "haven," "survey," "dplyr," and "multcomp."

Exhibit 12.4.A displays code that calculates the unweighted count, prevalence (mean), and totals for past month alcohol use (variable ALCMON) by sex for a single year and pooled years of data.

Exhibit 12.4.A Descriptive Statistics for Mean Estimates in R

```
TWOYEAR_DATA <-read_sas("SAS DATASET PATH AND NAME",
col_select=c(QUESTID,VESTR_C,VEREP,ANALWT2_C,YEAR,IRSEX,ALCMON,MRJMDA
       YS,MRJMON,CIGMON,IRWRKSTAT18,CATAG18))

names(TWOYEAR_DATA)<-tolower(names(TWOYEAR_DATA))
TWOYEAR_DATA$analwt2_c_2yr <- TWOYEAR_DATA$analwt2_c/2

create_design <- function(data, weight_var) {
  svydesign(
    id = ~verep,
    strata = ~vestr_c,
    data = data,
    weights = as.formula(paste("~", weight_var)),
    nest = TRUE
  ) %>%
    update(
      one = 1,
      irsex = factor(irsex, levels = 1:2, labels = c("Male",
        "Female"))
}

# Original design for estimates by single year of data
design <- create_design(TWOYEAR_DATA, "analwt2_c")

# Updated design for combined year estimates (using updated weight)
combined_design <- create_design(TWOYEAR_DATA, "analwt2_c_2yr")
degf(design)

sum(weights(design , "sampling") != 0 )

svyby(~one , ~ year , design , unwtd.count )
svyby(~one , ~ irsex , combined_design , unwtd.count )
svyby(~one , ~ year+irsex , design , unwtd.count )
```

Exhibit 12.4.A Descriptive Statistics for Mean Estimates in R (continued)

```
svytotal(~one, combined_design) %>% round
svyby(~one, ~ year, design, FUN=svytotal) %>% round
svyby(~one , ~ irsex , combined_design , FUN=svytotal)
svyby(~one , ~ year+irsex , design , FUN=svytotal)

svymean(~alcmon, combined_design, deff = "replace") %>% round(2)
svyby(~alcmon, ~year, design, svymean, deff = "replace")
svyby(~alcmon, ~irsex, design, combined_svymean, deff = "replace")
svyby(~alcmon, ~year+irsex, design, svymean, deff = "replace")

svytotal(~alcmon, combined_design)
svyby(~alcmon, ~year, design, svytotal)
svyby(~alcmon, ~irsex, combined_design, svytotal)
svyby(~alcmon, ~year+irsex, design, svytotal)
```

Exhibit 12.4.B displays code that shows how to compute estimates corresponding to levels of a categorical variable.

Exhibit 12.4.B Descriptive Statistics for Proportion Estimates in R

```
svyby(~ one , ~ mrjmon, design=subset(design, mrjmon==1),
        unwtd.count)
svytotal(~ mrjmon, design=subset(design, mrjmon==1), na.rm = TRUE)
svytotal(~ mrjmdays , design=subset(design, mrjmon==1), na.rm = TRUE)
b=svymean(~ mrjmdays, design=subset(design, mrjmon==1), na.rm = TRUE)
coef(b)*100
SE(b)*100
```

Exhibit 12.4.C displays code that manually computes a 95 percent confidence interval using output from R Exhibit 12.4.A.

Exhibit 12.4.C Manual Coding for Calculation of Confidence Interval in R

```
T_QNTILE=qt(c(.975), df=50); T_QNTILE
#DF of 50 is used in this example based on for the 2014-2023 NSDUH Public
        Use Files. See Table 10.1 for appropriate DF to use.

prop=svyby(~alcmon, ~year, design.R, svymean); prop
wdomain=svyby(~one, ~year, design.R, svytotal); wdomain

# For YEAR1 pop
NUMBER.Y1=SE(prop)[1]/(coef(prop)[1]*(1-coef(prop)[1])); NUMBER.Y1
L.Y1=log(coef(prop)[1]/(1-coef(prop)[1])); L.Y1
A.Y1=L.Y1-T_QNTILE*NUMBER.Y1; A.Y1
B.Y1=L.Y1+T_QNTILE*NUMBER.Y1; B.Y1
```

Exhibit 12.4.C Manual Coding for Calculation of Confidence Interval in R (continued)

```
# For YEAR2 pop
NUMBER.Y2=SE(prop)[2]/(coef(prop)[2]*(1-coef(prop)[2])); NUMBER.Y2
L.Y2=log(coef(prop)[2]/(1-coef(prop)[2])); L.Y2
A.Y2=L.Y2-T_QNTILE*NUMBER.Y2; A.Y2
B.Y2=L.Y2+T_QNTILE*NUMBER.Y2; B.Y2

# PLOWER AND PUPPER ARE THE 95% CIS ASSOCIATED WITH prevalence
PLOWER.Y1=1/(1+exp(-A.Y1)); PLOWER.Y1 # for YEAR1 pop
PUPPER.Y1=1/(1+exp(-B.Y1)); PUPPER.Y1 # for YEAR1 pop
PLOWER.Y2=1/(1+exp(-A.Y2)); PLOWER.Y2 # for YEAR2 pop
PUPPER.Y2=1/(1+exp(-B.Y2)); PUPPER.Y2 # for YEAR2 pop

# TLOWER AND TUPPER ARE THE 95% CIS ASSOCIATED WITH estimated total N
TLOWER.Y1=coef(wdomain)[1]*PLOWER.Y1; TLOWER.Y1 # for YEAR1 pop
TUPPER.Y1=coef(wdomain)[1]*PUPPER.Y1; TUPPER.Y1 # for YEAR1 pop
TLOWER.Y2=coef(wdomain)[2]*PLOWER.Y2; TLOWER.Y2 # for YEAR2 pop
TUPPER.Y2=coef(wdomain)[2]*PUPPER.Y2; TUPPER.Y2 # for YEAR2 pop
```

Exhibit 12.4.D displays code that computes the SE of the total for fixed domains using the data produced by Exhibit 12.4.A. Because sex is a fixed domain, the SE of the totals would not be taken directly from the R output but rather would be computed using the alternative SE estimation method.

Exhibit 12.4.D Manual Coding for Calculation of Standard Errors in R

```
wdomain=svyby(~one, ~year+irsex, design, svytotal)

# SE of proportion estimate of Alcohol drinker by sex and year
Seprop=svyby(~alcmon, ~year+irsex, design, svymean )
combined=cbind(subset(wdomain, select=-c(se)), subset(Seprop,
select=c(se))) #combine two stats together
combined$SE.FixedDomain=combined$one*combined$se; combined

# Repeat for combined sex by year
wdomaintot=svyby(~one, ~year, design, svytotal )
Seproptot=svyby(~alcmon, ~year, design, svymean )
combinedtot=cbind(subset(wdomaintot, select=-c(se)),
        subset(Seproptot,
select=c(se))) #combine two stats together

combinedtot$SE.FixedDomain=combinedtot$one*combinedtot$se;
        combinedtot
```

Exhibit 12.4.E displays code that shows how to create a log-linear chi-square test of independence for a subpopulation defined by three or more levels of a categorical variable. This code calculates Shah's Wald *F* test to indicate whether cigarette use is associated with employment status.

Exhibit 12.4.E Chi-Square Test of Independence in R

```
#prepare data
design.R <-
      svydesign(
           id = ~ verep ,
strata = ~ vestr_c ,
           data = DATANAME_SINGLE ,
weights = ~ analwt2_c ,
           nest = TRUE
      )

design.R=update(design.R,
      one = 1,
      irwrkstat18=factor(irwrkstat18, levels=1:4, labels =
           c("full-time", "part-time","unemployed","all other
                  adults"))
      )

svyby(~ one, ~ cigmon , subset(design.R, catag18==1), unwtd.count)
svyby(~ one, ~ irwrkstat18 , subset(design.R, catag18==1),
        unwtd.count)
svyby(~ one, ~ irwrkstat18+cigmon, subset(design.R, catag18==1),
        unwtd.count)

svytable(~irwrkstat18+cigmon, subset(design.R, catag18==1),
        round=TRUE)
svytable(~irwrkstat18+cigmon, subset(design.R, catag18==1),) %>%
prop.table(1)

a=svyloglin(~irwrkstat18+cigmon, subset(design.R, catag18==1))
b=update(a,~.^2); regTermTest(b, ~irwrkstat18:cigmon)
```

Once an association is determined by a significance Wald *F* test, then pairwise testing, as shown in Exhibit 12.4.F, can be done to determine individual significance tests across levels of employment status.

Exhibit 12.4.F Pairwise Testing in R

```
design.R <-
      update(design.R,
           employed1=ifelse(irwrkstat18==‘full-time’, 1, 0),
           employed2=ifelse(irwrkstat18==‘part-time’, 1, 0),
           employed3=ifelse(irwrkstat18==‘unemployed’, 1, 0),
           employed4=ifelse(irwrkstat18==‘all other adults’, 1, 0),
           group1=factor(ifelse(irwrkstat18 %in% c("full-
                time","parttime"), 1, 0), levels=0:1, labels=c("No",
                "Yes")),
           group2=factor(ifelse(irwrkstat18 %in% c("full-
                time","unemployed"), 1, 0), levels=0:1,
                labels=c("No", "Yes")),
           group3=factor(ifelse(irwrkstat18 %in% c("full-time","all
                other adults"), 1, 0), levels=0:1, labels=c("No",
                "Yes")),
           group4=factor(ifelse(irwrkstat18 %in%
                c("parttime","unemployed"), 1, 0), levels=0:1,
                labels=c("No", "Yes")),
           group5=factor(ifelse(irwrkstat18 %in% c("part-time","all
                other adults"), 1, 0), levels=0:1, labels=c("No",
                "Yes")),
           group6=factor(ifelse(irwrkstat18 %in% c("unemployed","all
                other adults"), 1, 0), levels=0:1, labels=c("No",
                "Yes"))
      )

svyby(~ one, ~ group1, design.R, unwtd.count)
svyby(~ one, ~ group2, design.R, unwtd.count)
svyby(~ one, ~ group3, design.R, unwtd.count)
svyby(~ one, ~ group4, design.R, unwtd.count)
svyby(~ one, ~ group5, design.R, unwtd.count)
svyby(~ one, ~ group6, design.R, unwtd.count)

svyby(~ one, ~ group1, design.R, svytotal)
svyby(~ one, ~ group2, design.R, svytotal)
svyby(~ one, ~ group3, design.R, svytotal)
svyby(~ one, ~ group4, design.R, svytotal)
svyby(~ one, ~ group5, design.R, svytotal)
svyby(~ one, ~ group6, design.R, svytotal)

a=svyglm(cigmon ~irwrkstat18, subset(design.R, catag18==1))
pw = summary(glht(a, mcp(irwrkstat18="Tukey")))
summary(pw, test = adjusted("none"))
summary(pw, test = adjusted("bonf"))
```

Exhibit 12.4.G displays code that performs significance testing between comparable years of NSDUH data. The input dataset would need to include at least 2 years of NSDUH data, but R does not require the data to be sorted by nesting variables.

Exhibit 12.4.G Tests of Differences for Proportions in R

```
count(DATANAME, year)
count(DATANAME, irsex)

svyby(~alcmon, ~year, design, svymean)
svyttest(alcmon~year, design)

svyby(~alcmon, ~year, subset(design , irsex == "male"), svymean)
svyttest(alcmon ~ year , subset(design , irsex == "male"))

svyby(~alcmon, ~year, subset(design , irsex == "female"), svymean)
svyttest(alcmon ~ year , subset(design , irsex == "female"))
```

Exhibit 12.4.H displays code that applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1 (commented out in the example).

Exhibit 12.4.H Manual Application of Suppression Rule in R

```
prop=svyby(~alcmon, ~year+irsex, design, svymean, deff = "replace");
       prop
ndomain=svyby( ~ one , ~ year+irsex , design , unwtd.count ); ndomain

prop=cbind(prop, subset(ndomain, select=c(counts))); prop
prop$RSE=ifelse(prop$se > 0.0, prop$se/prop$alcmon, NA)
prop$RSELNP=ifelse(prop$alcmon > 0.0 & prop$alcmon<=0.5,
prop$RSE/abs(log(prop$alcmon)),
     ifelse(prop$alcmon >0.5 & prop$alcmon<1.0,
     prop$RSE*((prop$alcmon/(1-prop$alcmon))/(abs(log(1-
            prop$alcmon)))), NA))

prop$EffNsum=prop$counts/prop$Deff.alcmon

#Suppression rule for proportion estimates: if suprule=1 then
       supress;
#do not if suprule=NA
prop$suprule=ifelse(prop$alcmon < 0.00005 | prop$alcmon> 0.99995 |
prop$RSELNP > 0.175 | prop$EffNsum < 68 | prop$counts <100, 1, NA);
       prop

#Use for Suppression rule for means (i.e., averages, not proportion)
#prop$suprule=ifelse((prop$RSE>0.5|prop$counts<10), 1, NA); prop
```

12.5 SPSS

The SPSS examples in this section provide guidance on how to use various programming options to produce estimates for NSDUH.

Exhibit 12.5.A displays code that computes the descriptive statistics means, sums, SEs, population size (POPSIZE), and design effect (DEFF) by survey year.

Exhibit 12.5.A Descriptive Statistics for Mean Estimates in SPSS

```
* Encoding: UTF-8.
GET
 FILE=‘SPSS DATASET PATH AND NAME’.
DATASET NAME Dataset1 WINDOW=FRONT.
*Sort dataset by Nesting Variables.
SORT CASES BY VESTR_C(A) VEREP(A).
*Create Complex Sampling Plan necessary for estimating variance from a
complex sample.
* ID Nesting variables (VESTR_C and VEREP) and weight variable
(ANALWT2_C
- standard single-year, person-level analysis weight). Alternatively,
a created pooled weight could be used here to produce annual averages
based on combined years of data.
CSPLAN ANALYSIS
 /PLAN FILE=‘PATH AND NAME TO SAVE DESIGN FILE’
 /PLANVARS ANALYSISWEIGHT=ANALWT2_C
 /SRSESTIMATOR TYPE=WR
 /PRINT PLAN
 /DESIGN STRATA=VESTR_C CLUSTER=VEREP
 /ESTIMATOR TYPE=WR.
*Create capture tag to store estimates into a dataset.
DATASET DECLARE ALC_EST.
OMS
 /SELECT TABLES
 /IF COMMANDS=[‘CSDescriptives’] SUBTYPES=[‘Univariate Statistics’]
 /DESTINATION FORMAT=SAV NUMBERED=TableNumber_
 OUTFILE=ALC_EST VIEWER=YES
 /TAG=estimates.
```

Exhibit 12.5.A Descriptive Statistics for Mean Estimates in SPSS (continued)

```
*Calculate overall by year estimates first.
* Year variable, where YEAR1=1 & YEAR2=2. Alternatively, the year
variable could identify the combined years of data, i.e., YEAR1 and
YEAR2 = 1 & YEAR3 and YEAR4 = 2.
DATASET ACTIVATE Dataset1.
CSDESCRIPTIVES
/PLAN FILE=‘PATH AND NAME OF DESIGN FILE’
/SUMMARY VARIABLES=ALCMON
/SUBPOP TABLE=YEAR DISPLAY=SEPARATE
/MEAN
/SUM
/STATISTICS SE POPSIZE DEFF COUNT
/MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
*Calculate sex by year estimates second.
* Sex variable, where male=1 & female=2.
CSDESCRIPTIVES
/PLAN FILE=‘PATH AND NAME OF DESIGN FILE’
/SUMMARY VARIABLES=ALCMON
/SUBPOP TABLE=YEAR BY IRSEX DISPLAY=SEPARATE
/MEAN
/SUM
/STATISTICS SE POPSIZE DEFF COUNT
/MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
OMSEND TAG =estimates.
*Remove rows that are not relevant (collapsed across years).
DATASET ACTIVATE ALC_EST.
SELECT IF (not(TableNumber_=1)).
SELECT IF (not(TableNumber_=4)).
EXECUTE.
*Transform estimates into standard publication formats.
DO IF Var1="Mean".
 Compute Percent=Estimate*100.
 Compute sePercent=StandardError*100.
END IF.
DO IF Var1="Sum".
 Compute Total=Estimate/1000.
 Compute seTotal=StandardError/1000.
 Compute DesignEffect=$sysmis.
 Compute PopulationSize=$sysmis.
 Compute UnweightedCount=$sysmis.
END IF.
EXECUTE.
```

Exhibit 12.5.B displays code that computes the SE of the total for fixed domains. Because sex is a fixed domain, the SE of the totals would not be taken directly from the SPSS output but rather would be computed using the alternative SE estimation method.

Exhibit 12.5.B Manual Coding for Calculation of Standard Errors in SPSS

```
*Recalculate the Standard Error of the Total since it is in a
controlled domain (Sex).
Compute seTOTAL=sePercent/100*PopulationSize/1000.
EXECUTE.
FORMATS Percent(F8.1).
FORMATS sePercent(F8.2).
FORMATS Total(COMMA8.0).
FORMATS seTotal(COMMA8.0).
FORMATS PopulationSize(COMMA8.0).
EXECUTE.
```

Exhibit 12.5.C displays code that applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1 (commented out in the example).

Exhibit 12.5.C Manual Application of Suppression Rule in SPSS

```
*SPSS stores totals and percentages within 2 different records, so
collapse to have all estimates on one row.
DATASET DECLARE ALC_EST2.
AGGREGATE
 /outfile=‘ALC_EST2’
 /BREAK= TableNumber_
 /Nsum PopSize DeffMean Percent sePercent Total seTotal
=sum(UnweightedCount PopulationSize DesignEffect Percent sePercent
Total seTotal).
EXECUTE.
FORMATS Percent(F8.1).
FORMATS sePercent(F8.2).
FORMATS Total(COMMA8.0).
FORMATS seTotal(COMMA8.0).
FORMATS POPSIZE(COMMA8.0).
EXECUTE.
*Apply Suppression Criteria.
COMPUTE mean=Percent/100.
COMPUTE semean=sePercent/100.
*Calculate Relative Standard Error (RSE).
DO IF (mean>0).
 COMPUTE RSE=semean/mean.
END IF.
* CALCULATE THE RELATIVE STANDARD ERROR OF NATURAL LOG P.
DO IF (mean GT 0 AND mean LE .5).
 COMPUTE RSELNP=RSE/ABS(LN(mean)).
END IF.
DO IF (mean GT .5 and mean LE 1.0).
 COMPUTE RSELNP=RSE*mean*(1-mean)/ABS(LN(1-mean)).
```

Exhibit 12.5.C Manual Application of Suppression Rule in SPSS (continued)

```
END IF.
*Calculate the Effective Sample Size.
COMPUTE EFFNSUM=NSUM/DEFFMEAN.
EXECUTE.
*SUPPRESSION RULE FOR PREVALENCE ESTIMATES.
DO IF (MEAN LT 0.00005 OR MEAN GT 0.99995 OR RSELNP GT 0.175 OR
EFFNSUM < 68 OR NSUM <100).
 COMPUTE SUPRULE=1.
END IF.
*SUPPRESSION RULE FOR MEANS NOT BOUNDED BY 0 AND 1, I.E. AVERAGES
(COMMENTED OUT FOR THIS EXAMPLE).
*IF (RSE GT 0.5 OR NSUM < 10).
 *COMPUTE SUPRULE=1.
*END IF.
EXECUTE.
```

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.).

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Center for Behavioral Health Statistics and Quality. (2022). *2023 National Survey on Drug Use and Health (NSDUH): Field interviewer manual* (Section 3.2.1). https://www.samhsa.gov/data/report/nsduh-2023-field-interviewer-manual

Center for Behavioral Health Statistics and Quality. (2024a). *2022 National Survey on Drug Use and Health (NSDUH) methodological resource book, Section 10: Editing and imputation report* (Chapters 2 and 3). https://www.samhsa.gov/data/report/nsduh-2022-editing-and-imputation-report

Center for Behavioral Health Statistics and Quality. (2024b). *2022 National Survey on Drug Use and Health (NSDUH) methodological resource book, Section 11: Person-level sampling weight calibration.* https://www.samhsa.gov/data/report/nsduh-2022-person-level-sampling-weight-calibration-report

Center for Behavioral Health Statistics and Quality. (2024c). *2022 National Survey on Drug Use and Health (NSDUH) methodological resource book, Section 13: Statistical inference report* (Chapters 6, 7, 8; Exhibits A.2.6, A.2.10). https://www.samhsa.gov/data/report/nsduh-2022-statistical-inference-report

Center for Behavioral Health Statistics and Quality. (2024d). *2023 companion infographic report: Results from the 2021, 2022, and 2023 National Surveys on Drug Use and Health* (SAMHSA Publication No. PEP24-07-020). https://www.samhsa.gov/data/report/2021-2022-2023-nsduh-infographic

Center for Behavioral Health Statistics and Quality. (2024e). *2023 National Survey on Drug Use and Health (NSDUH) methodological resource book, Section 2: Sample design report* (Chapter 2). https://www.samhsa.gov/data/report/nsduh-2023-sample-design-report

Center for Behavioral Health Statistics and Quality. (2024f). *2023 National Survey on Drug Use and Health (NSDUH): Methodological summary and definitions* (Sections 2.2, 2.3.4, 3.2.2, 3.2.3.2, 3.3.1). https://www.samhsa.gov/data/report/2023-methodological-summary-and-definitions

Center for Behavioral Health Statistics and Quality. (2024g). *2023 National Survey on Drug Use and Health public use file codebook* (Section 3). https://www.samhsa.gov/data/data-we-collect/nsduh/datafiles

IBM Corp. (2017). *IBM SPSS® statistics for windows*.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

RTI International. (2018). *SUDAAN®, Release 11.0.3* [computer software].

SAS Institute Inc. (2017). *SAS/STAT® software: Release 14.1* [computer software].

StataCorp LP. (2017). *Stata® statistical software: Release 14* [computer software].

Acknowledgments

This National Survey on Drug Use and Health (NSDUH) report was prepared by the Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality, and by RTI International. Work by RTI was performed under contract number 75S20322C00001. Marlon Daniel served as government project officer and as contracting officer representative, and David Hunter served as the RTI project director.

Contributors at SAMHSA included Samantha Brown, Xiaoting Qin, Shawn Linman, Jennifer Hoenig, and Tenecia Smith.

Contributors at RTI included Ana Saravia, Lauren Klein Warren, Teresa Davis, and Chelsy Thompson.

Appendix A: Examples of Key Indicators Measured in NSDUH

Table A.1 Variables Used in the 2023 Companion Infographic Report

Variable Name	Variable Label	Public Use File Codebook Section
Drug Use		
CIGMON	RC-CIGARETTES - PAST MONTH USE	RECODED DRUG USE
NICVAPMON	RC-NICOTINE VAPING - PAST MONTH USE	RECODED DRUG USE
ALCMON	RC-ALCOHOL - PAST MONTH USE	RECODED DRUG USE
MRJYR	RC-MARIJUANA - PAST YEAR USE	RECODED DRUG USE
COCYR	RC-COCAINE - PAST YEAR USE	RECODED DRUG USE
HALLUCYR	RC-HALLUCINOGENS - PAST YEAR USE	RECODED DRUG USE
METHAMYR	RC-METHAMPHETAMINE - PAST YEAR USE	RECODED DRUG USE
STMNMYR	RC-STIMULANTS - PAST YEAR MISUSE	RECODED DRUG USE
OPINMYR	RC-OPIOIDS - PAST YEAR MISUSE	RECODED DRUG USE
CNSNMYR	RC-CENTRAL NERVOUS SYSTEM STIMULANTS - PAST YEAR MISUSE	RECODED DRUG USE
ILLYR	RC-ANY ILLICIT DRUG - PAST YEAR USE	RECODED DRUG USE
ILLEMYR	RC-ILLICIT DRUG OTHER THAN MARIJUANA - PAST YEAR	RECODED DRUG USE
BNGDRKMON	RC-BINGE ALCOHOL USE PAST 30 DAYS	RECODED DRUG USE
HVYDRKMON	RC-HEAVY ALCOHOL USE PAST 30 DAYS	RECODED DRUG USE
Substance Use Disorder		
IRPYUD5ALC	ALCOHOL USE DISORDER IN THE PAST YEAR - IMP REV	IMPUTED SUBSTANCE USE DISORDER
UD5OPIANY	RC-OPIOID USE DISORDER - PAST YEAR USERS	RECODED SUBSTANCE USE DISORDER
UD5ILLANY	RC-DRUG USE DISORDER - PAST YEAR USERS	RECODED SUBSTANCE USE DISORDER
UD5ILALANY	RC-DRUG OR ALCOHOL USE DISORDER - PAST YEAR USERS	RECODED SUBSTANCE USE DISORDER
Alcohol and Drug Treatment		
SUTRTPY	RC-RCVD SUBSTANCE USE TREATMENT - PAST YEAR	RECODED ALCOHOL AND DRUG TREATMENT
Youth Experiences		
YUSUITHKYR	RC-YOUTH SERIOUSLY THOUGHT ABOUT KILLING SELF PAST YEAR	RECODED YOUTH EXPERIENCES
YUSUIPLNYR	RC-YOUTH MADE PLANS TO KILL SELF IN PAST YEAR	RECODED YOUTH EXPERIENCES
YUSUITRYYR	RC-YOUTH ATTEMPTED TO KILL SELF IN PAST YEAR	RECODED YOUTH EXPERIENCES

Table A.1 Variables Used in the 2023 Companion Infographic Report (continued)

Variable Name	Variable Label	Public Use File Codebook Section
Adult Mental Health		
IRSUICTHNK	ADULT SERIOUSLY THOUGHT ABOUT KILLING SELF PST YR - IMP REV	IMPUTED ADULT MENTAL HEALTH
IRSUIPLANYR	ADULT MADE PLANS TO KILL SELF IN PST YR - IMP REV	IMPUTED ADULT MENTAL HEALTH
IRSUITRYYR	ADULT ATTEMPTED TO KILL SELF IN PST YR - IMP REV	IMPUTED ADULT MENTAL HEALTH
SMIPY	RC-IMP SMI IND (1/0) BASED ON PREDICTED SMI PROB PY	RECODED ADULT MENTAL HEALTH
AMIPY	RC-IMP AMI IND (1/0) BASED ON PREDICTED SMI PROB PY	RECODED ADULT MENTAL HEALTH
AMISUD5ANYC	RC-IMP AMI AND SUB USE DISORDER - PAST YEAR-DSM-5 - ANY(1/0)	RECODED ADULT MENTAL HEALTH
Adult Depression		
IRAMDEYR	ADULT: PAST YEAR MAJOR DEPRESSIVE EPISODE (MDE) - IMP REV	IMPUTED ADULT DEPRESSION
Adolescent Depression		
YMDEYR	RC-YOUTH: PAST YEAR MAJOR DEPRESSIVE EPISODE (MDE)	RECODED ADOLESCENT DEPRESSION
YMSUD5YANY	RC-YOUTH: PY MDE AND SUB USE DISORDER-DSM-5 - ANY	RECODED ADOLESCENT DEPRESSION
Mental Health Services Utilization		
MHTRTPY	RC-RCVD MENTAL HEALTH TREATMENT - PAST YEAR	RECODED MENTAL HEALTH SERVICES UTILIZATION
Emerging Issues		
CASUPROB2	RC-PERCEIVED EVER HAD DRUG OR ALCOHOL USE PROBLEM	RECODED EMERGING ISSUES
RCVYSUBPRB	RC-PERCEIVED RECOVERY FROM DRUG OR ALCOHOL USE PROBLEM	RECODED EMERGING ISSUES
CAMHPROB2	RC-PERCEIVED EVER HAD A MENTAL HEALTH ISSUE	RECODED EMERGING ISSUES
RCVYMHPRB	RC-PERCEIVED RECOVERY FROM MENTAL HEALTH ISSUE	RECODED EMERGING ISSUES
IMFYR	RC-ILLEGALLY MADE FENTANYL - PAST YEAR USE	RECODED EMERGING ISSUES
FPIMFNMYR	RC-FENTANYL PRODUCTS INCLUDING IMF - PAST YEAR MISUSE	RECODED EMERGING ISSUES

Note: This table presents variables used in the 2023 Companion Infographic Report (Center for Behavioral Health Statistics and Quality, 2024d), which is available at https://www.samhsa.gov/data/report/2021-2022-2023-nsduh-infographic.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2021-2023.

Appendix B: NSDUH Resource Table

Table B.1 Links to NSDUH Resources

Title	Description	Link
NSDUH Frequently Asked Questions	Provide responses to frequently asked questions when getting started with NSDUH data.	https://www.samhsa.gov/data/faq-nsduh
NSDUH National Releases	Provide nationally representative data on the use of tobacco, alcohol, and illicit drugs; substance use disorders; receipt of substance use treatment; mental health issues; and the use of mental health services among the civilian, noninstitutionalized population aged 12 or older in the United States.	https://www.samhsa.gov/data/nsduh/national-releases
NSDUH Public Use File Codebooks	Provide documentation for the NSDUH public use data files.	https://www.samhsa.gov/data/data-we-collect/nsduh/datafiles
NSDUH State Data Tables and Reports	Provide reports, tables, and maps for states based on small area estimation.	https://www.samhsa.gov/data/nsduh/state-reports
NSDUH Variable Crosswalk Charts	Show the comparability of variables between years to allow for statistical analyses over several years of data.	https://www.samhsa.gov/data/data-we-collect/nsduh/datafiles
2022 and 2023 Methodological Resource Books	Provide context for NSDUH estimates and include materials used to conduct the survey, documentation on how the survey was conducted, and documentation on how the data were processed. New items are added to the collection as they become available, so documentation published in a previous Methodological Resource Book may not yet be available for the latest year. However, there are generally few changes from year to year, and a report from a previous year may have relevant information.	https://www.samhsa.gov/data/report/nsduh-2022-methodological-resource-book-mrb https://www.samhsa.gov/data/report/nsduh-2023-methodological-resource-book-mrb
2023 Companion Infographic Report	Presents a high-level visual representation of key NSDUH outcomes.	https://www.samhsa.gov/data/report/2021-2022-2023-nsduh-infographic
2023 Detailed Tables	Provide detailed national estimates from NSDUH including comprehensive data on substance use, mental health, and treatment in the United States. Appendix A contains a glossary of key definitions.	https://www.samhsa.gov/data/report/2023-nsduh-detailed-tables
2023 Methodological Summary and Definitions	Accompanies the annual detailed tables and national report and includes overall methodology, key definitions for measures and terms used in NSDUH reports and tables, and selected analyses of measures to aid with interpretation.	https://www.samhsa.gov/data/report/2023-methodological-summary-and-definitions
2023 NSDUH Questionnaire	Contains the English in-person and web interview questionnaires.	https://www.samhsa.gov/data/report/nsduh-2023-questionnaire

SAMHSA
Substance Abuse and Mental Health
Services Administration

Released 2025

SAMHSA's mission is to lead public health and service delivery efforts that promote mental health, prevent substance misuse, and provide treatments and supports to foster recovery while ensuring equitable access and better outcomes.

1-877-SAMHSA-7 (1-877-726-4727) | 1-800-487-4889 (TDD) | www.samhsa.gov