



# 2024 National Survey on Drug Use and Health (NSDUH) Methodological Resource Book

## Section 13: Statistical Inference Report

# 2024 National Survey on Drug Use and Health (NSDUH) Methodological Resource Book

## Section 13: Statistical Inference Report

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). Elizabeth Crane served as contracting officer representative, and Carlos Graham served as assistant contracting officer representative.

### Recommended Citation

Center for Behavioral Health Statistics and Quality. (2026). *2024 National Survey on Drug Use and Health (NSDUH) methodological resource book, Section 13: Statistical inference report*. <https://www.samhsa.gov/data/report/nsduh-2024-methodological-resource-book-mrb>

### Originating Office

Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, 5600 Fishers Lane, Rockville, MD 20857. Released 2026.

### 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

This publication is in the public domain and may be reproduced or copied without permission from SAMHSA. Citation of the source is appreciated. However, this publication may not be reproduced or distributed for a fee without the specific, written authorization of the Office of Communications, SAMHSA, HHS.

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 2026

## Contents

<b>1.</b>	<b><u>Introduction</u></b>	<b>1</b>
<b>2.</b>	<b><u>Background</u></b>	<b>4</b>
2.1	<u>Sample Design</u>	5
2.2	<u>Changes to Questionnaire Content and Survey Methodology</u>	8
2.2.1	<u>Suppression Rule</u>	8
2.2.2	<u>2024 Questionnaire Changes</u>	8
<b>3.</b>	<b><u>Prevalence Estimates</u></b>	<b>13</b>
3.1	<u>Measures That Use the Break-Off Weight</u>	14
<b>4.</b>	<b><u>Missingness</u></b>	<b>17</b>
4.1	<u>Potential Estimation Bias Due to Missingness</u>	17
4.2	<u>Variance Estimation in the Presence of Missingness</u>	20
<b>5.</b>	<b><u>Sampling Error</u></b>	<b>21</b>
<b>6.</b>	<b><u>Degrees of Freedom</u></b>	<b>26</b>
6.1	<u>Background</u>	26
6.2	<u>Degrees of Freedom Used in Key NSDUH Analyses</u>	29
<b>7.</b>	<b><u>Statistical Significance of Differences</u></b>	<b>31</b>
7.1	<u>Comparing Prevalence Estimates between Years</u>	31
7.1.1	<u>Example of Comparing Prevalence Estimates between Years</u>	34
7.1.2	<u>Example of Comparing Prevalence Estimates between Years in Excel</u>	35
7.2	<u>Comparing Prevalence Estimates in Categorical Subgroups</u>	36
7.2.1	<u>Testing among Levels of Categorical Subgroups</u>	36
7.2.2	<u>Testing among Levels of Categorical Subgroups: Age Adjustment</u>	37
7.2.3	<u>Significance Testing among Categorical Subdomains in Pooled Data from Multiple Years</u>	37
7.2.4	<u>Testing among a Subdomain and the Overall Population</u>	38
7.3	<u>Comparing Prevalence Estimates to Identify Linear Trends</u>	38
7.4	<u>Impact of Rounding in Interpreting Testing Results</u>	39

<b>8.</b>	<b><u>Confidence Intervals</u></b> .....	<b>40</b>
8.1	<u>Example of Calculating Confidence Intervals Using Published Prevalence Estimates and Standard Errors</u> .....	42
8.2	<u>Example of Calculating Confidence Intervals in Excel Using Published Prevalence Estimates and Standard Errors</u> .....	42
8.3	<u>Example of Calculating Standard Errors Using Published Confidence Intervals</u> .....	43
8.4	<u>Example of Calculating Standard Errors in Excel Using Published Confidence Intervals</u> .....	44
<b>9.</b>	<b><u>Suppression of Estimates with Low Precision and Rules for Presentation of Estimates</u></b> .....	<b>46</b>
9.1	<u>Suppression Rules for Proportions</u> .....	47
9.2	<u>Suppression Rule for Estimated Totals, Means, and Sample Sizes</u> .....	48
9.3	<u>Rounding Rules and Format for Presentation of Certain Unsuppressed Estimates</u> .....	49
	<b><u>References</u></b> .....	<b>50</b>
	<b><u>Acknowledgments</u></b> .....	<b>52</b>
	<b><u>Appendix A: Documentation for Conducting Various Statistical Procedures: SUDAAN®, Stata®, SAS®, R, and SPSS Examples</u></b> .....	<b>53</b>
A.1	<u>Guide for Defining Options for Analyzing NSDUH Data</u> .....	55
A.1.1	<u>NSDUH Sample Design</u> .....	55
A.1.2	<u>Nesting Variables</u> .....	56
A.1.3	<u>Degrees of Freedom</u> .....	56
A.1.4	<u>Design Effect</u> .....	57
A.1.5	<u>Standard Errors</u> .....	57
A.1.6	<u>Suppression Rule</u> .....	57
A.1.7	<u>Statistical Tests of Differences between Years</u> .....	58
A.1.8	<u>Recoding and Missing Values</u> .....	59
A.1.9	<u>Confidence Intervals</u> .....	59
A.1.10	<u>Calculating Percentages for Categories</u> .....	59
A.1.11	<u>Testing between Overlapping Domains</u> .....	59
A.1.12	<u>Testing Independence of Two Variables when One Variable Has Three or More Levels</u> .....	60
A.1.13	<u>Testing of Linear Trends</u> .....	60

<a href="#">A.2</a>	<a href="#">SUDAAN Exhibits</a>	<a href="#">61</a>
<a href="#">A.3</a>	<a href="#">Stata Exhibits</a>	<a href="#">73</a>
<a href="#">A.4</a>	<a href="#">SAS Exhibits</a>	<a href="#">86</a>
<a href="#">A.5</a>	<a href="#">R Code Exhibits</a>	<a href="#">95</a>
<a href="#">A.6</a>	<a href="#">SPSS Exhibits</a>	<a href="#">109</a>

## Exhibits

2.1	<a href="#">2024 NSDUH Sample Selection with Hybrid ABS and FE Frame</a>	6
6.1	<a href="#">97.5th Percentiles of <math>t</math>-Distributions for Varying Degrees of Freedom</a>	27
A.2.1	<a href="#">SUDAAN DESCRIPT Procedure (Estimate Generation: Single Year and Pooled Years of Data)</a>	61
A.2.2	<a href="#">SAS Code Based on SUDAAN Output (Calculation of Standard Error of Totals for Fixed Domains)</a>	62
A.2.3	<a href="#">SAS Code Based on SUDAAN Output (Implementation of Suppression Rule)</a>	63
A.2.4	<a href="#">SUDAAN DESCRIPT Procedure (Tests of Differences)</a>	63
A.2.5	<a href="#">SAS Code Based on SUDAAN Output (Calculation of the <math>P</math> Value for the Test of Differences between Totals for Nonfixed Domains)</a>	64
A.2.6	<a href="#">SUDAAN DESCRIPT Procedure and SAS Code Based on SUDAAN Output (Covariance Matrix, Identification of Covariance Components, Calculation of Variances, and Calculation of the <math>P</math> Value for the Test of Differences between Totals for Fixed Domains)</a>	64
A.2.7	<a href="#">SAS Code (Recoding a Variable) and SUDAAN DESCRIPT Procedure (Estimate Generation with (1) Missing Values and (2) Using Subpopulation)</a>	67
A.2.8	<a href="#">SAS Code Based on SUDAAN Output (Calculating a 95 Percent Confidence Interval)</a>	68
A.2.9	<a href="#">SUDAAN DESCRIPT Procedure (Estimate Generation for Categorical Variable, i.e., Number of Days Used Substance in the Past Month among Past Month Users)</a>	68
A.2.10	<a href="#">SAS Code (Stacking a Dataset) and SUDAAN DESCRIPT Procedure (Test of Difference when Two Groups Overlap Using Stacked Data)</a>	69
A.2.11	<a href="#">SUDAAN CROSSTAB Procedure (Test for Independence Based on a Log-Linear Model)</a>	70
A.2.12	<a href="#">SUDAAN DESCRIPT Procedure (Pairwise Testing)</a>	71
A.2.13	<a href="#">SUDAAN DESCRIPT Procedure (Test of Linear Trends with DESCRIPT)</a>	71
A.3.1	<a href="#">Stata COMMANDS svy: mean and svy: total (Estimate Generation: Single Year and Pooled Years of Data)</a>	73
A.3.2	<a href="#">Stata Code (Calculation of Standard Error of Totals for Fixed Domains)</a>	75
A.3.3	<a href="#">Stata Code (Implementation of Suppression Rule)</a>	75
A.3.4	<a href="#">Stata COMMANDS svy: mean and svy: total (Tests of Differences)</a>	76
A.3.5	<a href="#">Stata Code (Calculation of the <math>P</math> Value for the Test of Differences between Totals for Nonfixed Domains)</a>	78
A.3.6	<a href="#">Stata COMMAND svy: mean (Covariance Matrix, Identification of Covariance Components, Calculation of Variances, Calculation of the <math>P</math> Value for the Test of Differences between Totals for Fixed Domains)</a>	78
A.3.7	<a href="#">Stata Code (Recoding a Variable, Estimate Generation with (1) Missing Values and (2) Using Subpopulation)</a>	80
A.3.8	<a href="#">Stata Code (Calculating a 95 Percent Confidence Interval for a Mean)</a>	80

<a href="#">A.3.9</a>	<a href="#">Stata Code (Estimate Generation for Categorical Variable; i.e., Number of Days Used Substance in the Past Month among Past Month Users)</a>	<a href="#">81</a>
<a href="#">A.3.10</a>	<a href="#">Stata Code (Test of Difference when Two Groups Overlap Using Stacked Data)</a>	<a href="#">81</a>
<a href="#">A.3.11</a>	<a href="#">Stata Code (Test for Independence Based on a Log-Linear Model)</a>	<a href="#">82</a>
<a href="#">A.3.12</a>	<a href="#">Stata Code (Pairwise Testing)</a>	<a href="#">82</a>
<a href="#">A.3.13</a>	<a href="#">Stata Code (Test of Linear Trends Across Years)</a>	<a href="#">84</a>
<a href="#">A.4.1</a>	<a href="#">SAS SURVEYMEANS Procedure (Estimate Generation: Single Year and Pooled Years of Data)</a>	<a href="#">86</a>
<a href="#">A.4.2</a>	<a href="#">SAS Code Based on SAS Output (Calculation of Standard Error of Totals for Fixed Domains)</a>	<a href="#">86</a>
<a href="#">A.4.3</a>	<a href="#">SAS Code Based on SAS Output (Implementation of Suppression Rule)</a>	<a href="#">87</a>
<a href="#">A.4.4</a>	<a href="#">SAS Code (Tests of Differences)</a>	<a href="#">87</a>
<a href="#">A.4.5</a>	<a href="#">SAS Code (Calculation of the <i>P</i> Value for the Test of Differences between Totals for Nonfixed Domains)</a>	<a href="#">88</a>
<a href="#">A.4.6</a>	<a href="#">SAS Code (Covariance Matrix, Calculations of Variances, and Calculation of the <i>P</i> Value for the Test of Differences between Totals for Fixed Domains)</a>	<a href="#">88</a>
<a href="#">A.4.7</a>	<a href="#">SAS Code (Recoding a Variable, Estimate Generation with (1) Missing Values and (2) Using Subpopulation)</a>	<a href="#">90</a>
<a href="#">A.4.8</a>	<a href="#">SAS Code (Calculates a Confidence Interval for Alcohol Drinker Prevalence and Estimated Totals Produced in Exhibit A.4.1)</a>	<a href="#">91</a>
<a href="#">A.4.9</a>	<a href="#">SAS (Estimate Generation for Categorical Variable; i.e., Number of Days Used Substance in the Past Month among Past Month Users)</a>	<a href="#">92</a>
<a href="#">A.4.10</a>	<a href="#">SAS Code (Statistical Tests of Differences between Two Groups when the Two Groups Overlap)</a>	<a href="#">92</a>
<a href="#">A.4.11</a>	<a href="#">SAS Code (Tests of the Independence of the Prevalence Variable and Subgroup Variable)</a>	<a href="#">93</a>
<a href="#">A.4.12</a>	<a href="#">SAS Code (Pairwise Testing)</a>	<a href="#">93</a>
<a href="#">A.4.13</a>	<a href="#">SAS Code (Test of Linear Trends Across Years)</a>	<a href="#">94</a>
<a href="#">A.5.1</a>	<a href="#">R Code: svtotal and svymean (Estimate Generation: Single Year and Pooled Years of Data)</a>	<a href="#">95</a>
<a href="#">A.5.2</a>	<a href="#">R Code (Calculation of Standard Error of Totals for Fixed Domains)</a>	<a href="#">97</a>
<a href="#">A.5.3</a>	<a href="#">R Code (Implementation of Suppression Rule)</a>	<a href="#">98</a>
<a href="#">A.5.4</a>	<a href="#">R Code (Tests of Differences)</a>	<a href="#">99</a>
<a href="#">A.5.5</a>	<a href="#">R Code (Calculation of the <i>P</i> Value for the Test of Differences between Estimated Number Totals for Nonfixed Domains)</a>	<a href="#">99</a>
<a href="#">A.5.6</a>	<a href="#">R Code (Covariance Matrix, Calculations of Variances, and Calculation of the <i>P</i> Value for the Test of Differences between Totals for Fixed Domains)</a>	<a href="#">100</a>
<a href="#">A.5.7</a>	<a href="#">R Code (Estimate Generation with (1) Missing Values and (2) Using Subpopulation)</a>	<a href="#">103</a>
<a href="#">A.5.8</a>	<a href="#">R Code (Calculates a Confidence Interval for Alcohol Drinker Prevalence and Estimated Totals Produced in Exhibit A.5.1)</a>	<a href="#">104</a>

<a href="#">A.5.9</a>	<a href="#">R (Estimate Generation for Categorical Variable; i.e., Number of Days Used Substance in the Past Month among Past Month Users)</a> .....	104
<a href="#">A.5.10</a>	<a href="#">R Code (Statistical Tests of Differences between Two Groups when the Two Groups Overlap)</a> .....	105
<a href="#">A.5.11</a>	<a href="#">R Code (Tests of the Independence of the Prevalence Variable and Subgroup Variable)</a> .....	106
<a href="#">A.5.12</a>	<a href="#">R Code (Pairwise Testing)</a> .....	107
<a href="#">A.5.13</a>	<a href="#">R Code (Test of Linear Trends Across Years)</a> .....	108
<a href="#">A.6.1</a>	<a href="#">SPSS CSDESCRIPTIVES Procedure (Estimate Generation: Single Year and Pooled Years of Data)</a> .....	109
<a href="#">A.6.2</a>	<a href="#">SPSS Code Based on SPSS Output (Calculation of Standard Error of Totals for Fixed Domains)</a> .....	112
<a href="#">A.6.3</a>	<a href="#">SPSS Code (Implementation of Suppression Rule)</a> .....	112

## Tables

<a href="#">2.1</a>	<a href="#">2024 Target and Achieved Sample Allocation, by Age Group</a> .....	8
<a href="#">3.1</a>	<a href="#">Questionnaire Sections That Require Use of the Break-Off Analysis Weight and Imputed Variables That Use the Main Analysis Weight</a> .....	16
<a href="#">5.1</a>	<a href="#">Demographic and Geographic Domains Shown in the NSDUH National Reports and Detailed Tables Using the Alternative Standard Error Estimation Method for Calculating Standard Errors of the Estimated Number of People (Totals), 2024</a> .....	24
<a href="#">6.1</a>	<a href="#">Ninety-Five Percent Confidence Intervals for the Percentage of Past Month Users of Alcohol, Using Different Degrees of Freedom</a> .....	28
<a href="#">6.2</a>	<a href="#">Degrees of Freedom for Specific States per the 2014-2024 NSDUH Sample Design Based on the Restricted-Use Dataset</a> .....	28
<a href="#">6.3</a>	<a href="#">Key NSDUH Analyses and Degrees of Freedom for the Restricted-Use Data File and the Public Use Data File per the 2014-2024 NSDUH Sample Design</a> .....	30
<a href="#">9.1</a>	<a href="#">Summary of 2024 NSDUH Suppression Rules</a> .....	46
<a href="#">9.2</a>	<a href="#">Rounding Rules and Format for 2024 NSDUH Tables and Reports</a> .....	49
<a href="#">A.1</a>	<a href="#">Summary of SUDAAN, Stata, SAS, R, and SPSS Exhibits</a> .....	54
<a href="#">A.2</a>	<a href="#">SUDAAN Matrix Shell</a> .....	65
<a href="#">A.3</a>	<a href="#">Contrast Statements for Nonparametric Linear Trend Testing</a> .....	72
<a href="#">A.4</a>	<a href="#">Stata Matrix Shell</a> .....	78

## 1. Introduction

Statistical inference is a process of drawing a conclusion about the population based on survey data collected from a sample of that population. The target population for the 2024 National Survey on Drug Use and Health (NSDUH), conducted by RTI International,<sup>1</sup> was the U.S. civilian, noninstitutionalized population aged 12 or older (at the time of their interview). One example of conducting statistical inference would be using weighted responses from NSDUH to make a statement about the number of substance users<sup>2</sup> in the U.S. civilian, noninstitutionalized population, as well as the statistical uncertainty of that estimate.

Examples of the inferences made from the 2024 NSDUH data are presented in these key products:

- The *Key Substance Use and Mental Health Indicators in the United States: Results from the 2024 National Survey on Drug Use and Health* report (Center for Behavioral Health Statistics and Quality [CBHSQ], 2025f) is a national-level report focusing on estimates among people aged 12 years or older among the civilian, noninstitutionalized population in the United States. This report presents trends in estimates from 2021 to 2024.
- The *2024 Companion Infographic Report: Results from the 2021 to 2024 National Surveys on Drug Use and Health* (CBHSQ, 2025a) shows selected estimates from 2021 to 2024 among the population aged 12 or older.
- The Substance Abuse and Mental Health Services Administration (SAMHSA) has produced a series of reports using pooled data from the 2022-2024 NSDUHs that will examine associations in greater depth between selected demographic characteristics and substance use and mental health indicators. Reports in this series are available on the [NSDUH Data Collection](#) web page.
- *Results from the 2024 National Survey on Drug Use and Health: Detailed Tables* (CBHSQ, 2025h) are a comprehensive set of tables on substance use and mental health issues that include estimated numbers of people with a characteristic of interest (e.g., numbers of substance users, numbers of adults with mental illness), corresponding percentages, and standard errors of estimates. These tables show estimates for 2023 and 2024 and whether there was a statistically significant change between 2023 and 2024 when applicable. Multiyear tables are also included that show selected estimates for 2021-2024 and whether there are significant changes between estimates in 2024 and corresponding estimates in 2021-2023.

---

<sup>1</sup> RTI International is a trade name of Research Triangle Institute. RTI and the RTI logo are U.S. registered trademarks of Research Triangle Institute.

<sup>2</sup> NSDUH obtains information on the following 10 categories of drugs: marijuana, cocaine (including crack), heroin, hallucinogens, inhalants, and methamphetamine, as well as the misuse of prescription pain relievers, tranquilizers, stimulants, and sedatives. Estimates of “illicit drug use” reported from NSDUH reflect the use of drugs in any of these 10 categories.

Other supporting information relevant to estimates of substance use and mental health issues from the 2024 NSDUH can be found in the *2024 National Survey on Drug Use and Health (NSDUH): Methodological Summary and Definitions* report (CBHSQ, 2025c).

The 2024 NSDUH used multimode data collection, in which respondents could complete the survey via the web or in person. Methodological investigations led to the conclusion that estimates based on multimode data collection since 2021 are not comparable with estimates from 2020 or prior years. Chapter 6 in the 2021 Methodological Summary and Definitions report (CBHSQ, 2022) discusses these methodological investigations in greater detail. Therefore, 2022 through 2024 NSDUH estimates should not be compared with estimates prior to 2021 but can be compared with the 2021 NSDUH estimates.

Mode was included as a main effect in the person-level poststratification adjustment, with a 30 percent target for the web mode and a 70 percent target for the in-person mode to standardize the weighted proportions for each mode. This mode adjustment was applied to the NSDUH weights for 2021 onward to allow for comparisons between 2021 estimates and estimates from later years.<sup>3</sup> Without further adjustment to the weights, apparent increases in estimates between years could be due to a greater proportion of in-person interviews rather than real changes in the population. Stated another way, apparent increases in estimates could be partially due to the greater proportion of in-person respondents, not just to true changes in prevalence in the population. Similarly, decreases in prevalence may be partially obscured by the changes in proportions.

The purpose of this report is to describe the statistical inference procedures used to produce design-based estimates as presented in the 2024 national reports and tables, which are based on *restricted-use data*. The examples in this report are based on restricted-use NSDUH data. These examples utilize actual NSDUH data to aid users in understanding the concepts discussed in this report but are not directly linked to specific tables and reports. To emphasize key points for analyzing NSDUH data, certain sentences throughout this report appear in bold font.

The statistical procedures and information found in this report can also be generally applied to analyses based on the public use file; however, the results may be slightly different. Users of NSDUH's *public use data* may find inconsistencies in the variable names referenced in this report's Appendix A, the information presented in [Table 5.1](#), and other specific numbers presented in this report (i.e., degrees of freedom). For examples of statistical analyses using NSDUH public use data,<sup>4</sup> see Section 12 in the *2024 National Survey on Drug Use and Health: Public Use File Data Users' Guide* (CBHSQ, 2025e). For tables presenting estimates for selected measures based on the public use data, see Appendix F in the *2024 National Survey on Drug Use and Health Public Use File Codebook* (CBHSQ, 2025d).<sup>4</sup>

---

<sup>3</sup> The mode adjustment was applied to the 2021 NSDUH weights to create an updated person-level weight. For more information on the creation of the updated 2021 weights, investigations to update the 2021 weights, and the effects of the recalibrated 2021 weights on the 2021 estimates, see the 2022 Methodological Summary and Definitions report (CBHSQ, 2023a).

<sup>4</sup> As in previous years, CBHSQ will construct a public use data file for the 2024 NSDUH that will be available in late 2025 on the [NSDUH Data Collection](#) web page.

The remainder of this report is organized as follows: Chapter 2 provides background information concerning the survey design and questionnaire changes; Chapter 3 discusses the prevalence estimates and how they were calculated, including specifics on various topics presented in the detailed tables; Chapter 4 discusses how missing item responses of variables that are not imputed may lead to biased estimates; Chapter 5 discusses sampling errors and how they were calculated; Chapter 6 describes degrees of freedom and how they were used when comparing estimates; and Chapter 7 discusses how the statistical significance of differences between estimates was determined. Chapter 8 discusses confidence interval estimation, and Chapter 9 discusses the conditions under which estimates with low precision were suppressed. Appendix A contains examples that demonstrate how to conduct various statistical procedures documented within this report. Examples include using SUDAAN<sup>®</sup> Software for Statistical Analysis of Correlated Data (RTI International, 2020), Stata<sup>®</sup> (StataCorp LP, 2017), SAS<sup>®</sup> (SAS Institute Inc., 2023), R (R Core Team, 2024), and SPSS (IBM Corp, 2017). These examples are written in general terms (i.e., Year 1 instead of a specific year) to allow the example code to be more easily modified to fit the data users' analytic needs.

## 2. Background

The respondent universe for the National Survey on Drug Use and Health (NSDUH) is the civilian, noninstitutionalized population aged 12 or older residing within the 50 states and the District of Columbia.

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., people experiencing homelessness and 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 English or Spanish or are not physically or mentally capable of completing the interview are also excluded.

Survey data were either transmitted from field interviewers for in-person interviews or captured directly from the web-based data collection. These data were initially processed to create a raw data file that consists of one record for each usable interview, with no logical editing or other corrections. Data from this raw file underwent different types of processing to create final records and variables for analysis:

- removal of interviews with excessive amounts of missing data in the initial set of substance use questions;<sup>5</sup>
- editing of data for internal consistency; and
- statistical imputation to replace missing, inexact, or nonspecific values in key variables after data had been edited.

The final records and variables are referred to as *restricted-use data* in this report. For more information on the editing and imputation procedures, see the *2024 National Survey on Drug Use and Health (NSDUH) Methodological Resource Book, Section 10: Editing and Imputation Report* (Center for Behavioral Health Statistics and Quality [CBHSQ], forthcoming a).

The final respondent sample of 70,241 people for the 2024 NSDUH provides a sufficient sample to create population estimates for a broad range of ages and other demographic characteristics, geographic characteristics, and socioeconomic categories. Individual observations are weighted

---

<sup>5</sup> Interview records were eligible to be treated as final only if interview respondents provided data on lifetime use of cigarettes and at least 9 out of 13 of the other substances in the initial set of substance use questions.

so that the weighted sample represents the civilian, noninstitutionalized population aged 12 or older for the nation as a whole and for each state. **The person-level weights in NSDUH are calibrated by adjusting for nonresponse and poststratifying to known population estimates (or control totals) obtained from the U.S. Census Bureau and from the American Community Survey (ACS).**<sup>6</sup>

The rate of item nonresponse for web-based data collection is higher compared with that for in-person data collection due to respondents not completing the full survey (i.e., break-offs). Most missing data in usable interviews among adults were due to break-offs later in the survey. Missing data due to break-offs are potentially different from other missing data (i.e., responses of “don’t know” or “refused”). Treating them the same way in analyses could bias estimates, so break-off analysis weights were created for 2024 to account for items missing due to break-offs. For adults aged 18 or older, the break-off analysis weights were created by applying an additional adjustment to the main analysis weight. Because the number of break-offs among youths aged 12 to 17 was relatively small, the break-off analysis weight is equal to the main analysis weight for youths; that is, estimates for youths were the same regardless of which weight was used in the analysis. These break-off analysis weights were used for nonimputed measures from questions that were asked later in the survey. See Section 3.1 for more information on the specific measures included in the national reports and tables that use the break-off weight.

## 2.1 Sample Design

The NSDUH sample design uses stratified sampling, a selection method in which the population of interest is divided into subgroups (strata) before the research sample is chosen (Formplus Blog, 2023). When splitting the population into strata, naturally occurring divisors such as geographical location are used. Stratified sampling preserves the similarity within each stratum so that no subgroup is excluded from the final sample. However, differences between the population and the stratified sample may remain. Design-based weights based on the stratified, five-stage design of the study are used to adjust survey estimates to the target population. In addition, the stratified sample design is accounted for when computing variance. For more information, see the *2024 National Survey on Drug Use and Health (NSDUH) Methodological Resource Book, Section 2: Sample Design and Experience Report* (CBHSQ, 2025b). A state-based coordinated sample design was developed for the 2014-2024 NSDUHs, with an independent, multistage area probability sample within each state and the District of Columbia. As a result, states are viewed as the first level of stratification and as a variable for reporting estimates. Each state was further stratified into approximately equally populated state sampling

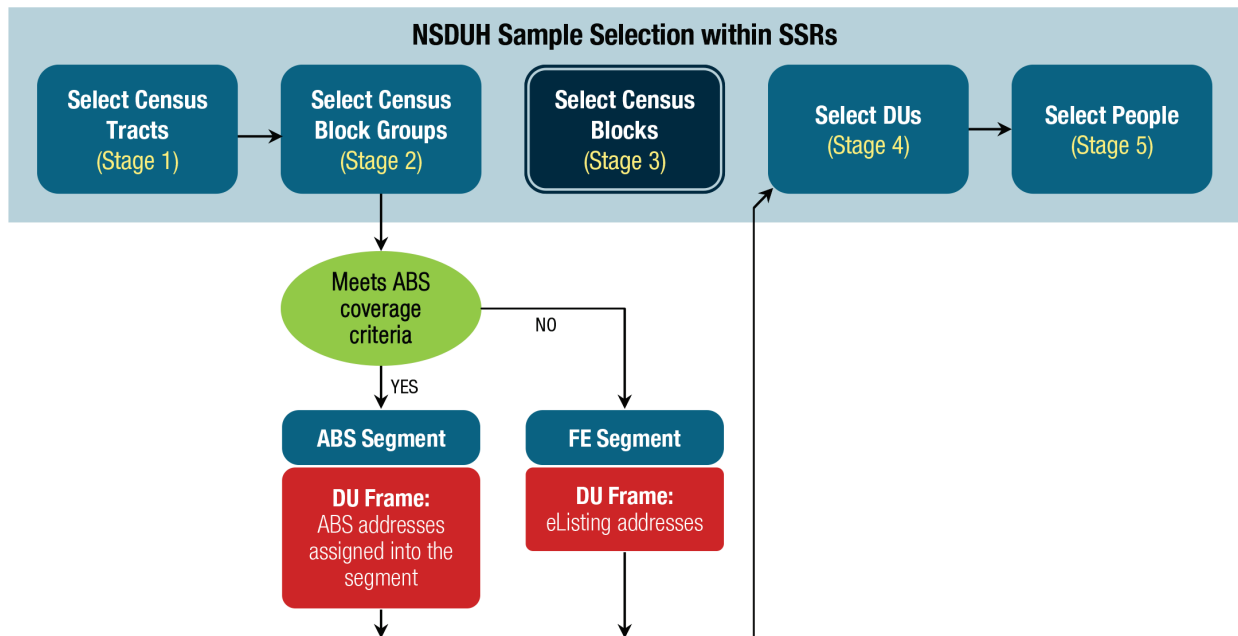
---

<sup>6</sup> For the 2024 weighting, the 2023 ACS data were used to create control totals for educational attainment. See the *2024 National Survey on Drug Use and Health (NSDUH) Methodological Resource Book, Section 11: Person-Level Sampling Weight Calibration* report (CBHSQ, forthcoming b) for more details.

regions (SSRs). SSRs were contiguous geographic areas designed to yield approximately the same number of interviews within a given state.<sup>2</sup>

The number of SSRs varied by state and was related to the state’s sample size. There was a total of 750 SSRs for 2024. There were then five stages of selection to create the multistage area probability sample (see [Exhibit 2.1](#)). First, census tracts were selected within each SSR (Stage 1), then census block groups were selected within those census tracts (Stage 2), and then smaller area segments (i.e., a collection of census blocks) were selected within the census block groups (Stage 3). Next, dwelling units (DUs) were selected within segments to receive a screener (Stage 4), and within each selected DU, up to two residents who were at least 12 years old were selected for the interview (Stage 5). If two eligible residents within the same DU were selected, they formed a within-DU pair.

### Exhibit 2.1 2024 NSDUH Sample Selection with Hybrid ABS and FE Frame



ABS = address-based sampling; DU = dwelling unit; FE = field enumeration; SSR = state sampling region.

A large reserve sample of area clusters was selected at the time that the 2014-2017 NSDUH sample was selected. This reserve sample was used to field the 2018-2024 NSDUHs. The coordinated sample design for 2014-2024 includes a 50 percent overlap in sampled areas within each successive 2-year period for 2014-2024 (e.g., 2014-2015, 2015-2016, 2016-2017). DUs that were not sampled the first year are eligible for selection the following year. There is no planned overlap of sampled residents and no longitudinal follow-up for individuals. However, people may be selected in consecutive years if they move and their new residence is selected. As seen in [Exhibit 2.1](#), the selection of smaller area segments within census block groups

<sup>2</sup> Sampling areas were defined using 2010 census geography. Counts of dwelling units and population totals were obtained from the 2010 decennial census data supplemented with revised population projections from [Claritas](#), a market research firm.

(Stage 3) was eliminated for the new portion of the 2024 sample. Selecting the DU samples from larger geographic areas was expected to increase precision. The 2014-2024 NSDUH sample design provides a sufficient number of completed interviews to support both state and national estimates.

For the 2024 NSDUH, a hybrid field enumeration and address-based sampling (ABS) approach was used to construct DU frames within sampled areas. ABS refers to the sampling of residential addresses from a list based on the U.S. Postal Service's Computerized Delivery Sequence file. Census block groups were evaluated using a set of ABS coverage criteria. If the census block group met all coverage criteria,<sup>8</sup> the ABS frame was used. If the census block group failed one or more coverage criteria, field enumeration was used to construct the DU frame (see [Exhibit 2.1](#)). To improve efficiencies and data quality, an electronic listing (eListing) application has been used since 2023 was used to enumerate DUs in field enumeration segments and to locate sampled DUs during data collection.

The 2014-2024 NSDUH sample design provides sufficient sample sizes to support state and national estimates. For the 2024 NSDUH, the target sample size for the largest 12 states was between 1,500 and 4,560 completed interviews and approximately 960 interviews in each of the remaining 37 states and the District of Columbia.<sup>9</sup> This cost-efficient sample design allocates completed interviews (and associated sample) to the largest 12 states approximately proportional to the size of the civilian, noninstitutionalized population aged 12 or older in these states. In the remaining states, a minimum sample size is selected in order to support reliable state estimates using either direct methods (by pooling multiple years of data<sup>10</sup>) or small area estimation.<sup>11</sup> Population projections based on the 2010 census and data from the 2006-2010 ACS were used to construct the sampling frame for the 2014-2024 NSDUHs. For detailed information on the 2024 sample design, see the 2024 Sample Design and Experience Report (CBHSQ, 2025b).

[Table 2.1](#) provides the target and achieved sample allocations for the 2024 NSDUH. Adolescents aged 12 to 17 years and young adults aged 18 to 25 years were oversampled. See the 2024 Sample Design and Experience Report (CBHSQ, 2025b) for more details on the sample design.

---

<sup>8</sup> See the 2024 Sample Design and Experience Report (CBHSQ, 2025b) for specifics about the ABS coverage criteria.

<sup>9</sup> For Hawaii, the sample was designed to yield a minimum of 200 completed interviews in Kauai County, Hawaii, over a 3-year period. To achieve this goal while maintaining precision at the state level, the annual sample in Hawaii consists of 67 completed interviews in Kauai County and 900 completed interviews in the remainder of the state, for a total of 967 completed interviews each year.

<sup>10</sup> Because of methodological changes, 2021 and later NSDUH data should never be pooled with NSDUH data from prior years for analyses.

<sup>11</sup> Small area estimation is a hierarchical Bayes modeling technique used to make state-level estimates for measures related to substance use and mental health. For details, see the "[2022-2023 National Survey on Drug Use and Health: Guide to State Tables and Summary of Small Area Estimation Methodology.](#)"

**Table 2.1 2024 Target and Achieved Sample Allocation, by Age Group**

Sample	12 to 17	18 to 25	26 or Older	26 to 34	35 to 49	50 or Older
Target	16,877 (25%)	16,877 (25%)	33,753 (50%)	10,126 (15%)	13,501 (20%)	10,126 (15%)
Achieved	13,985 (20%)	16,744 (24%)	39,512 (56%)	11,278 (16%)	15,379 (22%)	12,855 (18%)

NOTE: Percentages of the total sample are shown in parentheses.

NOTE: Achieved sample sizes are based on the reported age in the interview.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2024.

## 2.2 Changes to Questionnaire Content and Survey Methodology

NSDUH has undergone changes over the years to improve the quality of its data and to address the changing needs of policymakers and researchers with regard to substance use and mental health issues. These changes include updates to the questionnaire, data collection methods, and survey methodology. The next two subsections describe the 2024-specific changes to the questionnaire and changes in the suppression rule.

### 2.2.1 Suppression Rule

Beginning with SAMHSA publications that include 2024 NSDUH data, direct estimates from NSDUH are defined as unreliable based on the following:

- prevalence (for proportion estimates),
- relative standard error (defined as the standard error divided by the estimate), and
- sample size.

The suppression criteria for the prevalence rates were changed for data products using 2024 NSDUH data, primarily for greater simplicity. The new suppression criteria for 2024 were also applied to estimates from 2021 to 2023 presented in the *Results from the 2024 National Survey on Drug Use and Health: Detailed Tables* (CBHSQ, 2025h). Consequently, some estimates from 2021 to 2023 that were suppressed in prior years may be published in reports and tables for the 2024 NSDUH. See Chapter 9 for additional information on the 2024 suppression rule.

### 2.2.2 2024 Questionnaire Changes

NSDUH was conducted using multimode data collection in 2024, and the two forms of the questionnaire for web or in-person administration were kept as identical as possible. Hence, all notable changes to the questionnaire described in this section were made to both forms.

There were a large number of changes for the 2024 NSDUH, including dropping the coronavirus disease 2019 (COVID-19) section and all COVID-19-related questions. The following additional questions were removed from the questionnaire:

- Core Demographics
  - Removed questions about whether respondents were members of a reserve component currently serving full time in an active-duty status.

- Health
  - Removed questions about sexually transmitted diseases in the past 12 months.
- Youth Experiences<sup>12</sup>
  - Removed the questions about how adolescents’ friends would feel about them using certain substances.
  - Removed the question about whether adolescents participated in a program or meeting in the past year for their own or a family member’s substance use.
  - Removed the question about whether adolescents participated in pregnancy or sexually transmitted disease prevention programs in the past year.
  - Removed the question about whether adolescents had been exposed to any substance use prevention messages outside of school in the past year.
  - Removed the question about frequency of religious service attendance.
- Emerging Issues
  - Removed questions about marijuana vaping.
- Back-End Demographics
  - Removed the question about the state in which the respondent lived 1 year ago.
- Employment
  - Removed the question about workplace drug and alcohol use policies and access to workplace education and support.
  - Removed the questions about workplace drug and alcohol testing and respondent preferences for workplace testing.
- Household Roster
  - Removed the question about specific relationships of siblings in the household.

The following new questions were added to the 2024 questionnaire:

- Marijuana
  - Added a question about lifetime marijuana vaping to replace the question that was dropped from the emerging issues section.

---

<sup>12</sup> The question about whether adolescents spoke with at least one parent in the past year about the dangers of substance use was removed from the 2024 questionnaire. It has been added back to the 2025 questionnaire.

- Hallucinogens
  - Added a question asking about the most recent use of psilocybin and subsequent consistency check questions.
- Pain Relievers Screener
  - Added a series of five “OTHER, Specify” questions for any use of other prescription pain relievers to capture details on whether respondents used opioid or nonopioid pain relievers.
- Alcohol and Drug Treatment
  - Added a follow-up question for respondents who reported inpatient or outpatient treatment but did not report any substances for which they received treatment.
  - Added a question about the use of overdose reversal medicine in the past year.
- Youth Experiences
  - Added questions about the General Anxiety Disorder Scale (GAD-7) for adolescents.
- Mental Health
  - Added questions about the GAD-7 for adults.
- Back-End Demographics
  - Added a question to capture information on the state or U.S. territory in which respondents were born.

There were other changes to the questionnaire as well. Some may have an effect on comparability with specific previous estimates. See Chapter 3 of the *2024 National Survey on Drug Use and Health (NSDUH): Methodological Summary and Definitions* report (CBHSQ, 2025c) for the impact on specific measures.

The 2024 questionnaire included the following wording changes to existing items:

- Alcohol
  - Revised the introduction to update and simplify the definition of a drink.
  - Added help text to the introduction to show an updated and simplified list of example beverages.
- Cocaine
  - Revised the phrase “cocaine, in any form” to “any form of cocaine” to be more consistent with questions in other sections.

- Heroin
  - Revised the introduction to include example forms of heroin and ways of use.
- Inhalants
  - Revised the phrase “for kicks or to get high” to “for fun or to get high.” Questions about the last use of inhalants in the prior substance use section later in the questionnaire also had this revision.
- Youth Experiences
  - Revised the question about parental limits on time watching television to include parental limits on time with tablets, smartphones, computers, or video games in addition to television.
  - Revised the question about how the respondent’s parents would feel about the respondents smoking one or more packs of cigarettes a day to ask about smoking cigarettes every day.
  - Revised the question asking how adolescent respondents feel about their peers smoking one or more packs of cigarettes a day to ask about their peers smoking cigarettes every day. An introductory phrase was also added to this question.
- Alcohol and Drug Treatment<sup>13</sup>
  - Revised the introductory text about inpatient and outpatient alcohol and drug treatment to mention “any of the locations below” for response choices for specific locations.
  - Added text to response choices for each inpatient and outpatient alcohol and drug location to indicate that the treatment was “for alcohol or drug treatment.”
  - Revised the heading in the grid format for respondents to report the locations where they received inpatient treatment from “Stayed overnight or longer for alcohol or drug treatment in...” to “LOCATION.”
  - Revised the heading in the grid format for respondents to report the locations where they received outpatient treatment from “Received outpatient alcohol or drug treatment at...” to “LOCATION.”
- Mental Health Services Utilization<sup>13</sup>
  - Revised the introductory text about inpatient and outpatient mental health treatment to mention “any of the locations below” for response choices for specific locations.
  - Added text to response choices for each inpatient and outpatient mental health treatment location to indicate that the treatment was “for mental health, emotions, or behavior.”

---

<sup>13</sup> Changes to the alcohol and drug treatment and mental health service utilization sections were presented to respondents starting in June 2024.

- Revised the heading in the grid format for respondents to report the locations where they received inpatient treatment from “Stayed overnight or longer for treatment in...” to “LOCATION.”
- Revised the heading in the grid format for respondents to report the locations where they received outpatient treatment from “Received outpatient treatment at...” to “LOCATION.”

The 2024 questionnaire also included changes to existing items other than wording changes:

- Health
  - Updated the question about current pregnancy to be asked of female respondents aged 12 to 50. Before 2024, the question was asked of female respondents aged 12 to 44.
  - Updated the question about duration of the current pregnancy to ask about pregnancy length in weeks rather than months.
- Social Environment; Youth Experiences
  - For questions on attitudes about substance use in both the social environment section and the youth experiences section, updated the three-point scale ranging from “Neither approve nor disapprove” to “Strongly disapprove” to a five-point scale ranging from “Strongly disapprove” to “Strongly approve.”
- Back-End Demographics
  - Updated residency questions to be asked of respondents who were born in a U.S. territory in addition to those who were born in the United States.

Descriptions of changes to the 2024 NSDUH questionnaires can be found in the [2024 questionnaire specifications](#).

### 3. Prevalence Estimates

A prevalence estimate is an estimate of the **proportion of a population who has a specific characteristic in a given time period**. National prevalence estimates are reported as percentages and as the number of people with a specific characteristic in *Results from the 2024 National Survey on Drug Use and Health: Detailed Tables* (Center for Behavioral Health Statistics and Quality [CBHSQ], 2025h). **For example, the estimated percentage of people aged 12 or older who used marijuana in the past month is a prevalence estimate.**

For the national prevalence estimates, percentages are rounded to the nearest tenth of a percent, and population totals are presented in thousands. Unsuppressed rounded prevalence estimates of 0.0 are displayed as <0.1, and unsuppressed rounded standard errors of prevalence estimates of 0.0 are displayed as <0.01. Unsuppressed rounded numbers in thousands estimates of 0 and unsuppressed rounded standard errors of numbers in thousands estimates of 0 are displayed as <1. See Chapter 9 for the National Survey on Drug Use and Health (NSDUH) suppression rules. If an estimate is exactly a 0 value, corresponding to no respondents in the sample, the percentage and the number in thousands will be suppressed under the NSDUH suppression rule because of insufficient precision.

To illustrate the computation of prevalence estimates for characteristics of interest (such as substance use), let  $\hat{p}_d$  represent the prevalence estimate of interest for domain  $d$ . Then  $\hat{p}_d$  would be defined as the ratio

$$\hat{p}_d = \frac{\hat{Y}_d}{\hat{N}_d}, \quad (3.1)$$

where  $\hat{Y}_d = \sum_{i \in S} w_i \delta_i y_i$  represents the estimated number of people exhibiting the characteristic of interest in domain  $d$ ,  $\hat{N}_d = \sum_{i \in S} w_i \delta_i$  represents the estimated population total for domain  $d$ ,  $S$  represents the sample,  $w_i$  represents the analysis weight,  $\delta_i$  is defined as 1 if the  $i$ th sample unit is in domain  $d$  and is equal to 0 otherwise, and  $y_i$  is defined as 1 if the  $i$ th sample unit exhibits the characteristic of interest and is equal to 0 otherwise.

For certain populations of interest, sample sizes may not be adequate to support inferences using only 1 year of survey data. In these instances, estimates can be produced from annual averages based on combined data from 2 or more survey years (i.e., pooled data), although producing these estimates is possible for only years that have comparable data. Because of methodological changes, NSDUH data from 2021 and later years should never be pooled with data from 2020 and prior NSDUH years. **For pooled data, annual averages can be derived using the analysis weights divided by the number of years of combined data (see [Exhibit A.2.1](#), [Exhibit A.3.1](#), [Exhibit A.4.1](#), and [Exhibit A.5.1](#)).** For example, the weight variable would be divided by 2 for 2 years of combined data and by 4 for 4 years of combined data.

The published national estimates were computed using a multiprocedural package called SUDAAN<sup>®</sup> Software for Statistical Analysis of Correlated Data (RTI International, 2020).

**The final, nonresponse-adjusted, and poststratified analysis weights were used in SUDAAN to compute unbiased design-based estimates.** See the *2024 National Survey on Drug Use and Health (NSDUH) Methodological Resource Book, Section 11: Person-Level Sampling Weight Calibration* report for more information on the weights, including details about the main analysis weight and the break-off analysis weight (CBHSQ, forthcoming b). See Section 2.3.4 of the *2024 National Survey on Drug Use and Health (NSDUH): Methodological Summary and Definitions* report (CBHSQ, 2025c) for further details on how these weights were applied.

Appendix A contains examples that demonstrate how to compute the prevalence estimates using software packages including SUDAAN ([Exhibit A.2.1](#)), Stata<sup>®</sup> (StataCorp LP, 2017) ([Exhibit A.3.1](#)), SAS<sup>®</sup> (SAS Institute Inc., 2023) ([Exhibit A.4.1](#)), R (R Core Team, 2024) ([Exhibit A.5.1](#)), and SPSS (IBM Corp, 2017) ([Exhibit A.6.1](#)). For categorical measures, [Exhibit A.2.9](#), [Exhibit A.3.9](#), [Exhibit A.4.9](#), and [Exhibit A.5.9](#) demonstrate how to compute the prevalence estimates. No specific SPSS examples for computing prevalence estimates for categorical measures are included, but the same concepts apply.

The remainder of this chapter further discusses measures that require the use of the break-off weight when creating 2024 estimates. For additional information on measures not included here, such as the initiation measures or mental illness measures,<sup>14</sup> additional information can be found in other sources such as the 2024 Methodological Summary and Definitions report (CBHSQ, 2025c). For specific information about how variables are defined and created, see the data file codebooks.

### 3.1 Measures That Use the Break-Off Weight

As mentioned in Chapter 2, the use of web-based data collection in 2024 increased item nonresponse due to respondents not completing the full survey (i.e., break-offs). To reduce the potential bias that would arise from handling missing data due to break-offs in the same way that other missing data (i.e., responses of “don’t know” or “refused”) are handled in analyses, break-off analysis weights were created. These break-off analysis weights were used only for a small set of measures from questions that were asked later in the survey and that did not have missing values imputed. For 2024, the measures listed as follows are the ones that (1) were used in the 2024 Detailed Tables (CBHSQ, 2025h) or in the *Key Substance Use and Mental Health Indicators in the United States: Results from the 2024 National Survey on Drug Use and Health* report (CBHSQ, 2025f) and (2) used the break-off analysis weight. Any aggregate recodes based on these measures would also require use of the break-off weight. [Table 3.1](#) shows a list of questionnaire sections that require the use of the break-off weight starting with

---

<sup>14</sup> See Section 3.4.9 in the 2024 Methodological Summary and Definitions report (CBHSQ, 2025c) for more information on the creation of the mental illness measures. It is recommended that the mental illness measures derived from the model not be used when analyzing past year suicidal thoughts; past year major depressive episode (MDE); or other associated variables such as past year suicide attempts, past year suicide plans, past year medical treatment for suicide attempts, or lifetime MDE.

the mental health section and whether any variables in that section were imputed. For more information on these measures, see Appendix A in the 2024 Detailed Tables (CBHSQ, 2025h) or the data file codebooks.

- **Recovery from Substance Use Problems or Mental Health Issues:** Questions were included in the emerging issues section of the questionnaire. Respondents aged 18 or older first were asked whether they thought that they ever have had a problem with their own drug or alcohol use. If adult respondents answered “yes,” they were asked whether they considered themselves to be in recovery or to have recovered from their own problem with drug or alcohol use. These first two questions on recovery from a substance use problem were followed by a set of two similar questions asking adult respondents whether they have ever had a problem with their own mental health and, if so, whether they considered themselves to be in recovery or to have recovered from their own mental health problem.
- **College Enrollment:** Questions were included in the education section of the questionnaire. This measure was developed only for respondents aged 18 to 22 about current or upcoming enrollment in school and (if applicable) about whether respondents were full- or part-time students and the year of school they were or will be attending. Respondents were classified either as full-time college students or as some other status, which included respondents not enrolled in school, enrolled in college part time, enrolled in other grades either full time or part time, or enrolled with no other information available.
- **Underage Consumption of Alcohol:** Questions were included in the consumption of alcohol section of the questionnaire. Respondents who reported drinking at least one alcoholic beverage within the past 30 days and who were aged 12 to 20 were asked where they drank alcoholic beverages the last time they drank; if they were alone, with one other person, or with more than one person the last time they drank; and the source of the alcohol for their most recent alcohol use.
- **Drugs Used within 2 Hours of Alcohol:** Questions were included in the consumption of alcohol section of the questionnaire. Respondents who used alcohol in the past 30 days were asked if they used alcohol in combination with illicit drugs. The selected illicit drugs were marijuana, cocaine or crack, heroin, hallucinogens, inhalants, and methamphetamine.
- **Illegally Made Fentanyl Needle Use:** A question was included in the emerging issues section of the questionnaire. Respondents who reported having ever used, even once, illegally made fentanyl were asked if they had ever, even once, used a needle to inject illegally made fentanyl.

**Table 3.1 Questionnaire Sections That Require Use of the Break-Off Analysis Weight and Imputed Variables That Use the Main Analysis Weight**

Questionnaire Section	Variables That Require the Break-Off Analysis Weight <sup>1</sup>	Imputed Variables That Require the Main Analysis Weight <sup>2</sup>
Mental Health	Some	Kessler-6 variables on psychological distress; <sup>3</sup> WHODAS variables on impairment due to psychological distress; <sup>4</sup> serious thoughts of suicide, suicide plans, and suicide attempts in the past year; receipt of medical attention or a hospital stay because of a suicide attempt in the past year; and symptoms of generalized anxiety disorder <sup>5</sup>
Adult Depression	Some	Lifetime and past year MDE and past year MDE with severe impairment
Consumption of Alcohol	All	N/A
Emerging Issues	Some	Lifetime use and recency of use for kratom, synthetic marijuana, synthetic stimulants, vaping of flavoring, and illegally made fentanyl
Market Information for Marijuana	All	N/A
Back-End Demographics	Some	Immigrant status and immigrant age at entry to the United States
Education	All	N/A
Employment	Some	Employment status
Household Composition (Roster)	Some	Household size, number of people aged younger than 18, number of people aged 65 or older, other family in household, number of family members in household, and number of family members in household aged younger than 18
Proxy Information	All	N/A
Health Insurance	Some	Type of coverage (Medicare, Medicaid/CHIP, CHAMPUS, Private, Other)
Income	Some	Source of income (Social Security, Supplemental Security Income, food stamps, public assistance, welfare), months on welfare, personal income, and family income

MDE = major depressive episode; N/A = not applicable; WHODAS = World Health Organization Disability Assessment Schedule.

NOTE: The mental health and subsequent sections listed in this table are affected by interview break-offs. The break-off analysis weight should be used for analyses involving adult respondents and analyses involving the population aged 12 or older.

<sup>1</sup> A response of "All" indicates that the break-off analysis weight should be used for all variables from this section. A response of "Some" indicates that the break-off analysis weight is needed for use with variables in that section that were not imputed. Imputed variables in that section used the main analysis weight.

<sup>2</sup> Listed are specific variables or measures that were imputed. An overall composite measure may be imputed (e.g., past year MDE), but the individual source variables used to make a composite measure may not be imputed. Similarly, an overall composite measure itself may not be imputed (e.g., any mental illness, serious mental illness, poverty status), but all source variables were imputed. If no variables in a section were imputed, then the section is marked "N/A."

<sup>3</sup> See Section 3.4.9 in the *2024 National Survey on Drug Use and Health (NSDUH): Methodological Summary and Definitions* report (CBHSQ, 2025c) for a list of the Kessler-6 items.

<sup>4</sup> See Appendix A in the *Results from the 2024 National Survey on Drug Use and Health: Detailed Tables* (CBHSQ, 2025h) for a list of the WHODAS items included in the NSDUH questionnaire.

<sup>5</sup> See Section 3.4.11 in the 2024 Methodological Summary and Definitions report (CBHSQ, 2025c) for a list of the symptoms of generalized anxiety disorder.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2024.

## 4. Missingness

### 4.1 Potential Estimation Bias Due to Missingness

As is true for most large-scale surveys, the National Survey on Drug Use and Health (NSDUH) data have missing values for some questionnaire items. Missing values for the 2024 NSDUH can arise from respondents terminating the survey before reaching the end of the questionnaire (i.e., break-offs), not knowing how to answer a particular question (e.g., blank or “don’t know” responses), or refusing to answer a particular question. Missing values in NSDUH are handled either by (1) using statistical imputation where missing values are replaced with plausible values that are “donated” from a similar respondent with nonmissing data, and (2) excluding missing data from the analyses where unknown item responses are classified as missing values. Starting with the 2022 NSDUH, the zero-fill imputation method, where missing data are replaced with values equivalent to negative responses, was not applied for recoded analysis variables for the detailed tables.<sup>15</sup> However they are handled, missing data can cause biases in estimates, although bias is especially likely when missing values are assumed to be equivalent to negative responses. The effects of missing data on estimation as well as examples of how to determine the level of bias in estimation are discussed below.

In the 2024 NSDUH, many variables had missing item response values imputed, including but not limited to the main substance use variables, various demographic variables, and other topics listed in [Table 3.1](#). The imputation process treats the imputed value as a true response and therefore may underestimate the variance, but the difference is small enough to be considered ignorable. See the *2024 National Survey on Drug Use and Health (NSDUH) Methodological Resource Book, Section 10: Editing and Imputation Report* (Center for Behavioral Health Statistics and Quality [CBHSQ], forthcoming a) for further details on the imputation process and the evaluation on the impact of imputation on the variance.

When statistical imputation was not used to replace missing values with nonmissing values, NSDUH estimates were based on variables with some unknown data (e.g., blank, “don’t know,” “refused”). Generally, observations with missing values are excluded from standard NSDUH analyses. Excluding missing responses may have led to biased estimates in the *Results from the 2024 National Survey on Drug Use and Health: Detailed Tables* (CBHSQ, 2025h). [Exhibit A.2.7](#), [Exhibit A.3.7](#), [Exhibit A.4.7](#), and [Exhibit A.5.7](#), respectively, demonstrate how to compute prevalence estimates for variables with missing data using SUDAAN<sup>®</sup> Software for Statistical Analysis of Correlated Data (RTI International, 2020), Stata<sup>®</sup> (StataCorp LP, 2017), SAS<sup>®</sup> (SAS Institute Inc., 2023), and R (R Core Team, 2024).

As mentioned previously, item nonresponse for web-based data collection is higher than that for in-person data collection due to respondents not completing the full survey (i.e., break-offs).

---

<sup>15</sup> Because of the relatively large proportion of people who did not indicate the specific substance(s) for which they received treatment at some locations, a “substance unspecified” category was created. These respondents were grouped with the “no” category in the receipt of substance use treatment for specific substance recodes. See Section 3.4.8.3 in the *2024 National Survey on Drug Use and Health (NSDUH): Methodological Summary and Definitions* report (CBHSQ, 2025c) for more information.

Treating break-offs as equivalent to other missing data (i.e., responses of “don’t know” or “refused”) in analyses will not bias estimates when the probability of a break-off does not depend on the characteristics of respondents who broke off. However, for data based on later parts of the questionnaire, it was more likely that breaking off was related to the characteristics of respondents who broke off. To reduce the potential bias that would arise from handling missing data due to break-offs the same way that other missing data were handled, break-off analysis weights were created starting with the 2021 survey data based on questions from the adult mental health section and subsequent sections. To address potential nonresponse bias from sample members with less education being less likely to participate via the web, education was included in the poststratification adjustments for weighting. Additional special adjustments were made for the 2021 through 2024 NSDUH weights. See Section 2.3.4 in the *2024 National Survey on Drug Use and Health (NSDUH): Methodological Summary and Definitions* report (CBHSQ, 2025c) for more information on the weighting adjustments.

Investigations may be performed to look at the rates of missingness and the level of bias. Rates of missingness would be evaluated separately for each subpopulation to allow for detection of variations in missingness rates among different subpopulations. Recall from Formula 3.1 that prevalence estimates are defined as the proportions of the population who exhibit characteristics of interest.

Consider an analysis producing estimates for past year illicit drug use by probation status in the past year. The variable defining the characteristic of interest (e.g., illicit drug use) is referred to as the *analysis* variable, and the variable defining the domain of interest (e.g., probation status in the past year) is referred to as the *domain* variable. Suppose that the analysis variable has all its missing values imputed, but the domain variable does not employ the imputation of missing values. In such instances, the estimates from Formula 3.1 of  $\hat{N}_d$  and  $\hat{Y}_d$  may be negatively biased, and the  $\hat{p}_d$  estimates also may be biased. To see this, suppose that the domain variable has  $D$  levels, and define

$$\hat{N} = \sum_{d=1}^D \hat{N}_d + \hat{N}_m, \quad (4.1)$$

where  $\hat{N}$  = estimated population total,  $\hat{N}_d$  = estimated population total for domain  $d$ ,  $d = 1, 2, \dots, D$ , and  $\hat{N}_m$  = estimated population total corresponding to the missing values of the domain variable. Thus, if  $\hat{N}_m$  is positive (i.e., there are missing domain-variable responses), then at least one of the  $\hat{N}_d$  estimates will be negatively biased. The presence of negative bias in at least one of the  $\hat{Y}_d$  estimates can be similarly demonstrated if  $\hat{Y}_m$  is positive, where  $\hat{Y}_m$  = the estimated number of people exhibiting the characteristic of interest and corresponding to the missing values of the domain variable. **If either  $\hat{N}_m$  or  $\hat{Y}_m$  is positive, then  $\hat{p}_d$  may be biased by some unknown amount.**

Suppose instead that the domain variable has all its missing values imputed, but the analysis variable does not employ the imputation of missing values. In such instances, at least one of the  $\hat{N}_d$  estimates will be negatively biased. If all missing values for the analysis variable in the domain do not have the condition of interest,  $\hat{Y}_d$  would have no bias. Otherwise,  $\hat{Y}_d$  will be negatively biased. Thus,  $\hat{p}_d$  may be biased by some unknown amount. Likewise,  $\hat{p}_d$  may be biased when the domain and analysis variables do not employ the imputation of missing values. A negative bias is created with magnitude dependent on the percentage of respondents with missing data and on the magnitude of the estimate. Specifically, higher levels of nonresponse paired with high estimates induce a larger negative bias. A lower level of nonresponse paired with lower prevalence estimates induces a smaller negative bias. Intermediate combinations induce a moderate negative bias. In the 2024 Detailed Tables (CBHSQ, 2025h), **potential bias in the  $\hat{N}_d$ ,  $\hat{Y}_d$ , or  $\hat{p}_d$  estimates was not treated, although footnotes included on the tables provide detailed information about which estimates included or excluded missing values.**

The following example illustrates the approximate level of bias in analyses of the NSDUH data that exclude missing data. This example uses NSDUH data from the restricted-use file. First, consider past year illicit drug use estimates among adults aged 18 or older who reported being on probation or not in the past year, where unknown past year probation status is classified as a missing value. The population estimate of adults aged 18 or older was approximately 250,316,000. However, the subdomain population estimates for "on probation" and "not on probation" summed to approximately 249,634,000 due to excluding missing data, resulting in an estimate of  $\hat{N}_m = 682,000$  (approximately 0.3 percent of the total population). This number represents the estimated population not assigned to either domain. This negative bias can extend to various analysis variables, such as "Illicit Drugs." The total estimate of adults aged 18 or older who used illicit drugs in the past year was approximately 33,636,000. However, the estimate of adults aged 18 or older who used illicit drugs in the past year among the valid subdomains (where past year probation status was not missing) summed to 33,569,000, resulting in an estimate of  $\hat{Y}_m = 66,470$  (approximately 0.2 percent of the total population aged 18 or older who used illicit drugs in the past year). Because  $\hat{N}_m$  is positive and  $\hat{Y}_m$  is positive for the "Illicit Drugs" analysis variable, the prevalence estimates may be biased by some unknown amount across the two domains. The prevalence estimates of illicit drug use reported for adults aged 18 or older who were on probation or not on probation in the past year are 32.9 and 13.1 percent, respectively. By recoding the item missingness of the domain variable "Probation" as being on or not being on probation, the approximate range of possible bias values for each of these estimates is as follows: between -6.45 and 9.26 percent for on probation and between -0.06 and 0.06 percent for not on probation.

## 4.2 Variance Estimation in the Presence of Missingness

SUDAAN (RTI International, 2020) uses the number of strata<sup>16</sup> and number of primary sampling units (PSUs) in its variance calculations, even if there are some PSUs in which a variable is entirely missing for all sample members associated with that PSU. The rationale behind this approach is that, even if no sample members have nonmissing values in that PSU, there may be people in the target population who do.

To illustrate how this is operationalized in SUDAAN, consider the following example. Suppose someone tries to calculate the mean of some variable (say,  $X$ ), but there are missing values associated with variable  $X$ . SUDAAN then creates an internal subpopulation indicator variable (say,  $\delta$ ), where  $\delta = 1$  if variable  $X$  is not missing, and  $\delta = 0$  if variable  $X$  is missing. SUDAAN then internally calculates the mean and variance of variable  $X$  by using  $\delta X$ , assuming that the full sample mean is the same as the nonmissing sample mean.

For the variance estimator based on the Taylor series linearization approach, one of the terms in the variance estimator consists of the sum of squared deviations of PSU-level totals about their stratum-level means, divided by the number of PSUs in the stratum minus 1. Therefore, if SUDAAN encounters an incorrect number of PSUs within a stratum, then this term is incorrectly calculated. In addition, if there is only one PSU in a stratum, then the denominator for the variance term associated with that stratum becomes 0, which causes the overall variance estimate to return an error message in SUDAAN. **By including all PSUs in a stratum, whether or not the PSU has reported values, SUDAAN computes the variances appropriately. That is, PSUs with nothing but missing values for a variable should never be excluded from an input file. Thus, users are encouraged to use the full NSDUH dataset when running analyses in order to keep the complete data structure for variance estimation. Subsetting of the data to populations of interest should be done within SUDAAN (e.g., using SUDAAN's SUBPOPN statement).**

---

<sup>16</sup> See Chapter 6 of this report for more information on the number of strata for the 2024 NSDUH.

## 5. Sampling Error

In sampling, statistics from different samples will vary and can differ from the true population parameter; this difference is called *error*. Sampling error is the error caused by using statistics based on a sample instead of interviewing every person in the population. Standard errors (SEs) are commonly used to describe how much statistics differ from the true parameter due to sampling. This measure is incorporated in common statistical methods such as significance testing (see Chapter 7) and confidence intervals (see Chapter 8). **Like the prevalence estimates, all the variance estimates for prevalence (including those for prevalence based on annual averages from combined data) for the 2024 National Survey on Drug Use and Health (NSDUH) national data products were calculated using a method in SUDAAN® Software for Statistical Analysis of Correlated Data (RTI International, 2020) 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.

Because of the complex nature of the sampling design for NSDUH (specifically, the use of stratified cluster sampling), key nesting variables were created for use in SUDAAN to capture explicit stratification and to identify clustering. Each state sampling region (SSR) appears in a different variance estimation stratum every quarter. This method has the effect of assigning the regions to strata in a pseudo-random fashion while ensuring that each stratum consists of four SSRs from four different states.

A NSDUH replicate is one of two distinct and virtually independent unions of primary sampling units, each of which could be used to compute a nearly unbiased population estimate (of a state, division, region, or the entire United States in a quarter, a year, or a combination of years). Two variance replicates (VEREP) per year are defined within each variance stratum (VESTR). Each variance replicate consists of four segments, one for each quarter of data collection. One variance replicate consists of those segments that are “phasing out” or will not be used in the next survey year. The other variance replicate consists of those segments that are “phasing in” or will be fielded again the following year, except in 2024, when the “phasing in” segments will not be used again. A segment stays in the same VEREP for the 2 years it is in the sample. This simplifies computing SEs for estimates based on combined data from adjacent survey years.<sup>17</sup>

**Although the SEs of means and proportions can be calculated appropriately in SUDAAN using a Taylor series linearization approach, the actual SEs of estimates of totals may be smaller in situations where the domain size is poststratified to data from the U.S. Census Bureau or the American Community Survey.** Because of the potential for gains in precision, alternatives for estimating SEs of totals were implemented in all of the *Results from the 2024 National Survey on Drug Use and Health: Detailed Tables* (Center for Behavioral Health Statistics and Quality [CBHSQ], 2025h), where appropriate.

---

<sup>17</sup> Because of methodological changes, NSDUH data from 2021 and onward should never be pooled with NSDUH data from years prior to 2021 for analyses.

The SUDAAN software package is used to calculate direct estimates of  $\hat{Y}_d$  and  $\hat{N}_d$  using Formula 3.1 and can be used to estimate their respective SEs. A Taylor series approximation method implemented in SUDAAN provides estimates for  $\hat{p}_d$  and its SE.

When the domain size,  $\hat{N}_d$ , is assumed to be free of sampling error, Formula 5.1 below is an alternative to using SUDAAN to estimate the SE for the total number of persons with a characteristic of interest (e.g., substance users):

$$SE(\hat{Y}_d) = \hat{N}_d SE(\hat{p}_d) \quad (5.1)$$

This alternative SE estimation method is theoretically correct when the domain size estimates,  $\hat{N}_d$ , such as for the sex category, are fixed (i.e., among those domains forced to match their respective U.S. Census Bureau or American Community Survey population estimates through the weight calibration process). In these situations,  $\hat{N}_d$  is not subject to a sampling error induced by the NSDUH design. For more information, see the *2024 National Survey on Drug Use and Health Methodological Resource Book, Section 11: Person-Level Sampling Weight Calibration* report (CBHSQ, forthcoming b).

For an estimated number  $\hat{Y}_d$ , where the domain  $\hat{N}_d$  is nonfixed (i.e., where domain size estimates, such as for the race category of Asian, are not forced to match the U.S. Census Bureau or American Community Survey population estimates), this alternative SE estimation method still may provide a good approximation if it can be assumed that the sampling variation in  $\hat{N}_d$  is negligible relative to the sampling variation in  $\hat{p}_d$ . This is a reasonable assumption for many NSDUH analyses.

For various subsets of estimates, using this alternative SE estimation method where domain sizes are nonfixed yielded an underestimate of the variance of a total because  $\hat{N}_d$  was subject to considerable variation. Because of this underestimation, the alternative SE estimation method was not implemented when  $\hat{N}_d$  was nonfixed.

For the 2024 NSDUH, a “mixed-method” approach was used in tables to improve the accuracy of SEs for the estimated numbers of people and to better reflect the effects of poststratification on the variance of the total estimated numbers of people. **This approach assigns the method of SE calculation to domains within tables so that all estimates among a select set of domains with fixed  $\hat{N}_d$  were calculated using the alternative SE estimation method, and all other estimates were calculated directly in SUDAAN, regardless of other estimates within the same table.** The set of domains with a fixed  $\hat{N}_d$

was restricted to main effects and two-way interactions to maintain continuity between years.<sup>18</sup> Domains consisting of three-way interactions may be fixed in one year but not necessarily in preceding or subsequent years. The use of the mixed-method approach did not affect the SE estimates for the corresponding proportions presented in the same sets of tables because all SEs for means and proportions are calculated directly in SUDAAN. Appendix A contains SUDAAN, Stata® (StataCorp LP, 2017), SAS® (SAS Institute Inc., 2023), R (R Core Team, 2024), and SPSS (IBM Corp., 2017) examples that demonstrate how to compute SEs of proportions and both types of SEs of totals (see [Exhibit A.2.2](#), [Exhibit A.3.2](#), [Exhibit A.4.2](#), [Exhibit A.5.2](#), and [Exhibit A.6.2](#), respectively).

[Table 5.1](#) contains a list of domains used in the *Key Substance Use and Mental Health Indicators in the United States: Results from the 2024 National Survey on Drug Use and Health* (CBHSQ, 2025f) and the 2024 Detailed Tables (CBHSQ, 2025h) that employ the alternative SE estimation method for the restricted-use data file, including the main domains and the two-way interactions.<sup>19</sup> For domains not included in [Table 5.1](#), SEs for the estimates of totals are calculated directly in SUDAAN. To illustrate, consider a table presenting estimates of any mental illness (AMI) among adults aged 18 or older within the domains of age group, sex, Hispanic origin and race, and current employment. The estimated numbers of adults with AMI among the total population and age group (age group is the main effect), males and females (age group by sex interaction), and people who were Hispanic or not Hispanic (age group by Hispanic origin interaction) would use the alternative SE estimation method to calculate the SEs. The SEs for all other categories in this illustration of estimated numbers of people would be calculated directly in SUDAAN because NSDUH estimates in the national reports and detailed tables for racial groups are among people who were not Hispanic, unless noted otherwise. For example, the SEs by age group for White people were calculated directly in SUDAAN. The domain for White people is actually for White people who were not Hispanic and is a two-way interaction. Therefore, age group for White people is considered a three-way interaction, and the SEs by age group for White people were calculated directly in SUDAAN. Current employment is also not a fixed domain, and the SE of the estimated number of people would be calculated directly in SUDAAN.

In addition to the domains illustrated above, all four levels of education are treated as a fixed domain for adults aged 18 or older. Although not reported in the 2024 Key Substance Use and Mental Health Indicators report (CBHSQ, 2025f) and 2024 Detailed Tables (CBHSQ, 2025h), additional geographic interactions are also treated as domains with fixed  $\hat{N}_d$  for other NSDUH analyses. Similar to geographic region, geographic division, individual states, two-way interactions with state and sex, Hispanic origin, quarter, age group (12 to 17, 18 to 25, and 26

---

<sup>18</sup> Not all the race domains in [Table 5.1](#) are forced to fully match the U.S. Census Bureau population estimates due to models not converging. When this occurs, the sampling variation in  $\hat{N}_d$  for these domains is considered negligible. Therefore, the race domains are considered fixed, and the alternative method is applied the same way every year.

<sup>19</sup> See the standard error of the totals section in the *2024 National Survey on Drug Use and Health: Public Use File Data Users' Guide* for a list of domains that employ the alternative SE estimation method for the 2024 public use data file (CBHSQ, 2025e).

or older), and the two-way interaction between geographic region and age group are treated as domains with fixed  $\hat{N}_d$ , which would all employ the alternative SE estimation method.

Additionally, quarter is treated as a domain with fixed  $\hat{N}_d$ , as is the two-way interaction with state, sex, and age group.

**Table 5.1 Demographic and Geographic Domains Shown in the NSDUH National Reports and Detailed Tables Using the Alternative Standard Error Estimation Method for Calculating Standard Errors of the Estimated Number of People (Totals), 2024**

Main Effects	Two-Way Interactions <sup>1</sup>
<b>Age Group</b>	<b>Age Group × Sex</b>
12-17	(e.g., males aged 12 to 17)
18-25	-
26-34	<b>Hispanic Origin × Age Group</b>
35-49	(e.g., Hispanic or Latino people aged 18 to 25)
50-64	-
65 or Older	<b>Age Group × Geographic Region</b>
Collapsed Age Group Categories from Above <sup>2</sup>	(e.g., people aged 12 to 25 in the Northeast)
<b>Sex</b>	<b>Sex × Hispanic Origin</b>
Male	(e.g., not Hispanic or Latino males)
Female	-
<b>Hispanic Origin</b>	-
Hispanic or Latino	-
Not Hispanic or Latino	-
<b>Race<sup>3</sup></b>	<b>White, Not Hispanic or Latino</b>
<b>Geographic Region</b>	-
Northeast	-
Midwest	-
South	-
West	-
<b>Education (Aged 18 or Older)</b>	-
Less than High School	-
High School Graduate	-
Some College/Associate's Degree	-
College Graduate	-

NOTE: The alternative standard error (SE) estimation method for the estimated number of people (totals),

$SE(\hat{Y}_d) = \hat{N}_d SE(\hat{p}_d)$ , is applied when the domain size estimates,  $\hat{N}_d$ , are among those forced to match their respective U.S. Census Bureau or American Community Survey (ACS) population estimates through the weight calibration process.

NOTE: The alternative SE estimation method does not affect the SEs for the corresponding means and proportions. These latter SEs are calculated directly in SUDAAN® (RTI International, 2020), whereas the alternative SE estimation method is computed outside of SUDAAN using the formula provided in the first note.

(continued)

**Table 5.1 Demographic and Geographic Domains Shown in the NSDUH National Reports and Detailed Tables Using the Alternative Standard Error Estimation Method for Calculating Standard Errors of the Estimated Number of People (Totals), 2024 (continued)**

NOTE: This table shows only the domains and domain combinations used in the *Results from the 2024 National Survey on Drug Use and Health: Detailed Tables* (Center for Behavioral Health Statistics and Quality [CBHSQ], 2025h) and *Key Substance Use and Mental Health Indicators in the United States: Results from the 2024 National Survey on Drug Use and Health* report (CBHSQ, 2025f). Other domains and domain combinations (omitted here) also use this alternative SE estimation method, but they are not included in these specific reports or tables. For example, methodological studies or special requests often include a wider variety of domains and survey years. This variation requires the SE method to be assessed for each individual analysis. For a detailed list of domains for NSDUH forced to match their respective U.S. Census Bureau or ACS population estimates through the weight calibration process, see the *2024 National Survey on Drug Use and Health (NSDUH) Methodological Resource Book, Section 11: Person-Level Sampling Weight Calibration* report (CBHSQ, forthcoming b).

NOTE: The domains using the alternative SE estimation method for calculating the SE of the estimated number of people (total) are the same for both the main analysis weight and the break-off analysis weight.

- <sup>1</sup> Unless otherwise noted, the domains for the two-way interactions 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.
- <sup>2</sup> Main-effect age group categories shown in the table can be collapsed to form broader age group categories (e.g., 12 or older, 50 or older, 18 to 49, 26 to 49). Collapsed main-effect age group categories and two-way interactions with other main-effect demographic or geographic domains shown (e.g., males aged 50 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, 15 to 44), age groups finer than the main-effect age groups (e.g., 12 to 13, 18 to 20), or two-way interactions of these types of collapsed age categories with other main-effect domains (e.g., females aged 15 to 44) should not use the alternative SE estimation method.
- <sup>3</sup> Race is included as a main effect in this table for completeness; however, racial groups include all people within a given racial category, regardless of whether they are Hispanic or not Hispanic. In contrast, groups other than Hispanic in the national report tables and detailed tables are indented under the "Non-Hispanic" ethnicity row heading. For example, the domain for White people in the 2024 Detailed Tables (CBHSQ, 2025h) is actually non-Hispanic White people and is therefore a two-way interaction. Thus, any additional domains crossed with non-Hispanic White people (e.g., White people aged 18 to 25) represent three-way interactions not using the alternative SE estimation method.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2024.

## 6. Degrees of Freedom

### 6.1 Background

The degrees of freedom ( $df$ ) are needed to determine whether the observed difference between estimates is statistically significant. To make this determination, a test statistic and a probability level ( $p$  value) both need to be determined. The test statistic is computed from the sample data and represents a numerical summary of the difference between the estimates under consideration; it is a random variable that has a predetermined distribution (such as Student's  $t$ , chi-square, or  $F$ ). **The  $df$  characterize the amount of variation expected in the estimation of sampling error and are used in conjunction with the test statistic to determine probabilities and evaluate statistical significance.**

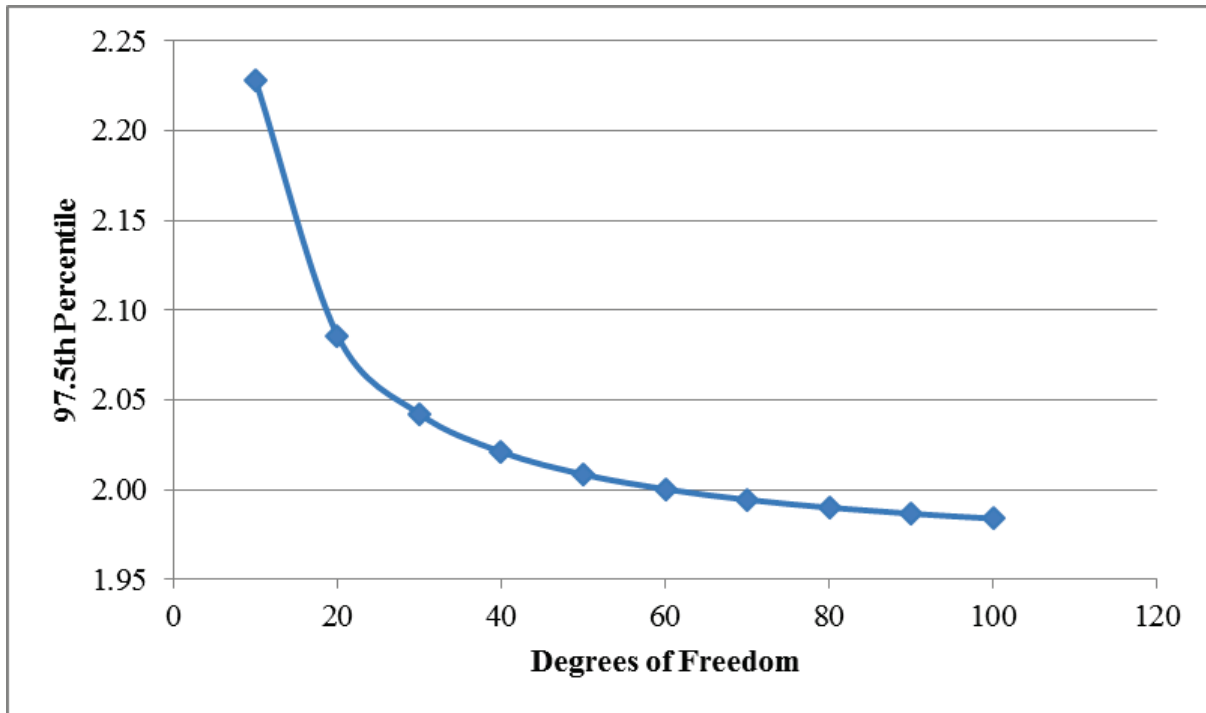
In statistics, the number of  $df$  refers to the number of independent units of information in a sample relevant to the estimation of a parameter or calculation of a statistic. In general, the  $df$  of a parameter estimate are equal to the number of independent observations that go into the estimate minus the number of other parameters that need to be estimated as an intermediate step. The  $df$  are also used to compute the confidence intervals (CIs) discussed in Chapter 8. The upper and lower limits of the CIs are defined by a constant value that is chosen to yield a level of confidence based on the  $df$ .

In practice, beyond a certain value, which  $df$  value is used has little impact. For example, the 97.5th percentile of the  $t$ -distribution is used in the National Survey on Drug Use and Health (NSDUH) to create 95 percent CIs and for two-sided hypothesis tests, and this does not change much once there are about 50  $df$ . Thus, results with 50  $df$  are similar to results with the 750  $df$  used for the 2014-2024 NSDUHs ([Exhibit 6.1](#)). In addition, [Table 6.1](#) shows the large sample 95 percent CI for a "typical" estimate for different  $df$ . The CIs are similar.

The  $df$  for analyses vary based on the geographic area of interest. In the 2014-2024 NSDUH design, sampling strata called state sampling regions (SSRs) were formed within each state. **Each SSR appears in a different variance stratum each quarter (i.e., variance strata are defined at the SSR and quarter levels). When rolled up to the national level, there are a total of 750 variance strata, which results in 750  $df$  for national estimates.** If an analysis involves individual states, the  $df$  are determined by the number of variance strata in which the state is included. The NSDUH design has five state sample size groups. Each of the smallest sample states is in 48 different variance strata (12 SSRs  $\times$  4 quarters = 48 variance strata); therefore, 48  $df$  are available for state estimates in these states. At the other extreme, the largest sample state, California, is in 144 variance strata (36 SSRs  $\times$  4

quarters = 144 variance strata) and therefore has 144 *df* for estimation. [Table 6.2](#) shows the *df* for specific states per the 2014-2024 NSDUH sample design.<sup>20</sup>

**Exhibit 6.1 97.5th Percentiles of *t*-Distributions for Varying Degrees of Freedom**



<sup>20</sup> Users of the 2024 public use file (Center for Behavioral Health Statistics and Quality [CBHSQ], 2025d) may find inconsistencies with the specific *df* presented in this report because the specific information referenced is based on the restricted-use dataset that was used to create the *Results from the 2024 National Survey on Drug Use and Health: Detailed Tables* (CBHSQ, 2025h) and the *Key Substance Use and Mental Health Indicators in the United States: Results from the 2024 National Survey on Drug Use and Health* report (CBHSQ, 2025f).

**Table 6.1** Ninety-Five Percent Confidence Intervals for the Percentage of Past Month Users of Alcohol, Using Different Degrees of Freedom

Degrees of Freedom	Critical Value of the <i>t</i> -Distribution	95% Confidence Interval	
		Lower Limit	Upper Limit
10	2.2281	50.00	51.53
20	2.0860	50.05	51.49
30	2.0423	50.07	51.47
40	2.0211	50.07	51.46
50	2.0086	50.08	51.46
60	2.0003	50.08	51.46
70	1.9944	50.08	51.45
80	1.9901	50.09	51.45
90	1.9867	50.09	51.45
100	1.9840	50.09	51.45
500	1.9647	50.09	51.44
750	1.9631	50.09	51.44
900	1.9626	50.09	51.44
1,800	1.9613	50.10	51.44

NOTE: The percentage of past month users of alcohol used to produce the data in this table is 50.77 percent, with a corresponding standard error of 0.34, both rounded to 2 decimal places.

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

**Table 6.2** Degrees of Freedom for Specific States per the 2014-2024 NSDUH Sample Design Based on the Restricted-Use Dataset

States	Sample Design Years	Degrees of Freedom
California	2014-2024	144
Florida, New York, and Texas	2014-2024	120
Illinois, Michigan, Ohio, and Pennsylvania	2014-2024	96
Georgia, New Jersey, North Carolina, and Virginia	2014-2024	60
Remaining 38 states and the District of Columbia	2014-2024	48

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2014-2024.

Appendix A contains examples that demonstrate how to define the *df* in SUDAAN® Software for Statistical Analysis of Correlated Data (RTI International, 2020), Stata® (StataCorp LP, 2017), SAS® (SAS Institute Inc., 2023), and R (R Core Team, 2024), which are used to compute CIs and perform significance testing.

Under the NSDUH sample design, for an analysis of a group of states, the *df* would be less than or equal to the sum of the *df* for each individual state due to overlap of strata. **Therefore, for subnational estimates, the specific number of *df* should be computed by counting**

**the unique values of the applicable VESTR variable (variance estimation [pseudo] stratum) for the particular geographic area of interest.** For these types of specific state analyses (or other subpopulations of interest), the *df* can be calculated outside SUDAAN and this value entered manually into SUDAAN for use in testing; otherwise, the *df* are computed using the entire dataset. Similar methods can be used to compute appropriate *df* for any geographic region comprising counties. Using this technique with the public use file will give similar, but not always exact, results. The technique of counting the number of unique values of VESTR can also be used to compute the number of *df* for subnational analyses based on combining survey data across years.<sup>21</sup> [Exhibit A.2.1](#), [Exhibit A.3.1](#), [Exhibit A.4.1](#), [Exhibit A.5.1](#), and [Exhibit A.6.1](#) can be adjusted to compute estimates based on pooled data.

## 6.2 Degrees of Freedom Used in Key NSDUH Analyses

The current practices for applying *df* to NSDUH data depend on the type of analyses. [Table 6.3](#) summarizes key types of NSDUH analyses and the *df* used for these analyses per the sample design. **The Key Substance Use and Mental Health Indicators report and detailed tables use the national *df* for the most current survey year (including census region and division and estimates for all years and pooled data when applicable<sup>21</sup>), with the exception of estimates for the mean age of first use (AFU) and the average number of days used.** The current year *df* are used because when conducting significance testing between estimates with different *df*, the lower *df* provide a more conservative test.<sup>22</sup> For all the currently analyzed years of NSDUH data, the current year's *df* have always been less than or equal to the previous years' *df*.

AFU and average number of days used estimates are treated differently because of the possibility of smaller sample sizes for these means (e.g., the sample sizes for AFU estimates are typically the number of past year initiates rather than the entire population). In that case, they often belong to fewer variance estimation strata. Based on the NSDUH suppression rules, the sample size threshold for suppression of a mean is 10, compared with 50 for a prevalence estimate.<sup>23</sup> Thus, it is possible for nonsuppressed estimates of averages to have smaller sample sizes than prevalence estimates. For example, the subpopulation for estimates of mean AFU includes only past year initiates of prescription drugs and lifetime users of other drugs, which could be small for drugs with low rates of use. **For all average estimates, including the mean AFU, the number of nonempty strata are used as the *df* in order to produce more conservative tests than are produced using the national *df*.**

Unlike the detailed tables, which use the national *df* for estimates by geographic subgroup (census region), special analyses and methodological reports follow the procedures described in Section 6.1 for this subgroup. The *df* used for key NSDUH analyses are summarized in

---

<sup>21</sup> Because of methodological changes, NSDUH data from 2021 and onward should never be pooled with NSDUH data from years prior to 2021 for analyses.

<sup>22</sup> Because of methodological changes, NSDUH data from 2021 and onward should never be tested against NSDUH data from years prior to 2021.

<sup>23</sup> The suppression criteria for the prevalence rates were changed in 2024 data products. See Section 9 for more details.

**Table 6.3. For NSDUH analyses that compare two geographic subpopulations (including those that compare subpopulations with the full population), the standard practice is to use the smaller of the two values for *df* to err on the side of being conservative.** For analyses where the subpopulation is not geographic in nature (e.g., members of a certain race or age category, past year users of a certain drug), the standard practice is to use the same *df* value that is used for analyses involving the whole population.

**Table 6.3 Key NSDUH Analyses and Degrees of Freedom for the Restricted-Use Data File and the Public Use Data File per the 2014-2024 NSDUH Sample Design**

Analyses	Sample Design Years	Degrees of Freedom for Restricted-Use (Public Use) Data File <sup>1</sup>
Special analyses involving the whole population or a nongeographic subpopulation <sup>2</sup>	2014-2024	750 (50)
Special analyses involving a single state	2014-2024	See <a href="#">Table 6.2</a>
Special analyses involving other geographic subpopulations <sup>3</sup>	Any	Count of the unique values of applicable VESTR variable (variance estimation [pseudo] stratum) for the particular geographic area of interest <sup>4</sup>
Detailed tables or Key Substance Use and Mental Health Indicators reports with estimates of averages, including mean age at first use	2014-2024	Number of nonempty <sup>2</sup> strata (for each estimate/subpopulation)
All other detailed tables and Key Substance Use and Mental Health Indicators reports (including those involving geographic subpopulations)	2014-2024	750 (50)

<sup>1</sup> The degrees of freedom shown first in this column are based on the restricted-use data files, and the degrees of freedom in parentheses are based on the public use data file. State is not available on the public use data file; thus, only information on the degrees of freedom based on the restricted-use data files is provided.

<sup>2</sup> 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.

<sup>3</sup> Some analyses capped the degrees of freedom at 900, regardless of year combinations across the study year groups. This rule is not consistently applied to all special analyses and reports.

<sup>4</sup> Users of the 2024 public use file (Center for Behavioral Health Statistics and Quality, 2025d) may find inconsistencies in the counts when comparing them with published data.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2014-2024.

## 7. Statistical Significance of Differences

Once the degrees of freedom ( $df$ ) have been determined as described in Chapter 6, various methods used to compare prevalence estimates can be used. This chapter describes the methods used to compare prevalence estimates, examples showing how to compute the comparison of estimates between years and within years among subdomains, and the impact of rounding in interpreting testing results.

Customarily, one way that observed difference between estimates is evaluated is in terms of its statistical significance. Although significance tests are often used to distinguish whether a difference is “real” or simply occurring due to sampling, it is important to note that (1) a “real” difference does not necessarily mean a policy-relevant difference, and (2) tests are based on probability and may give a false impression of certainty. Statistical significance is based on the size of the test statistic and its corresponding  $p$  value, which refers to the probability that a difference as large as that observed would occur because of random variability in the sample estimates if there were no differences in the population prevalence values being compared. The significance level is the chosen cutoff value for when a  $p$  value is considered small enough to call a difference “significant.” Generally, differences are reported as significant at the 0.05 or 0.01 levels.

Significance tests can only be conducted on differences between comparable prevalence estimates (i.e., estimates based on either a single year of data or combined years of comparable survey data) from comparable years of the National Survey on Drug Use and Health (NSDUH). **Data users should exercise caution when comparing estimates and confirm comparability of the measures and years in question. Estimates should never be compared when there is a trend break.** Estimates based on multimode data collection starting in 2021 are not comparable with estimates from 2020 or prior years; therefore, significance testing should not be conducted between estimates from 2021 and estimates from prior years. To compare 2021 estimates with estimates from 2022 and future years, data users must use the adjusted 2021 weight.

Significance tests can also be conducted on differences between prevalence estimates from combined years of comparable survey data. In NSDUH products, within-year tests were also conducted on differences between prevalence estimates for various populations (or subgroups) of interest using data from the 2024 survey. Tests comparing individual subgroups with the full population for demographics can also be conducted. See Section 7.2 for specifics on comparing among subgroups.

### 7.1 Comparing Prevalence Estimates between Years

When comparing prevalence estimates, the null hypothesis (no difference in the prevalence estimates) can be tested against the alternative hypothesis (there is a difference in the prevalence estimates) using the standard  $t$  test (with the appropriate  $df$ ) for the difference in proportions test, expressed as

$$t_{df} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\text{var}(\hat{p}_1) + \text{var}(\hat{p}_2) - 2\text{cov}(\hat{p}_1, \hat{p}_2)}}, \quad (7.1)$$

or

$$t_{df} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\text{var}(\hat{p}_1) + \text{var}(\hat{p}_2) - 2\rho(\hat{p}_1, \hat{p}_2)\text{SE}(\hat{p}_1)\text{SE}(\hat{p}_2)}}, \quad (7.2)$$

where in both formulas,  $df$  = the appropriate degrees of freedom,  $\hat{p}_1$  = the first prevalence estimate,  $\hat{p}_2$  = the second prevalence estimate,  $\text{var}(\hat{p}_1)$  = the variance of the first prevalence estimate, and  $\text{var}(\hat{p}_2)$  = the variance of the second prevalence estimate. In the first formula,  $\text{cov}(\hat{p}_1, \hat{p}_2)$  = covariance between  $\hat{p}_1$  and  $\hat{p}_2$ . In the second formula, the covariance between  $\hat{p}_1$  and  $\hat{p}_2$  is displayed as the product of the correlation between  $\hat{p}_1$  and  $\hat{p}_2$  and the standard errors (SEs) of  $\hat{p}_1$  and  $\hat{p}_2$ , where  $\rho(\hat{p}_1, \hat{p}_2)$  = the correlation between  $\hat{p}_1$  and  $\hat{p}_2$  and  $\text{SE}(\hat{p}_1)\text{SE}(\hat{p}_2)$  = the product of the SEs for  $\hat{p}_1$  and  $\hat{p}_2$  (i.e., the two formulas are equivalent; the first formula is defined in terms of the covariance, and the second is defined in terms of the correlations and SEs). Generally, the correlations between estimates in adjacent years are very small and positive; thus, ignoring the correlation in the second formula will usually result in a slightly more conservative test outcome, which is a test that is less likely to reject the null hypothesis that there is no difference in the two estimates. However, a negative correlation is possible and would result in a liberal test, which means it would be more likely to reject the null hypothesis that there is no difference in the two estimates. Additionally, the second (simplified) formula can be used in the case of two independent (i.e., uncorrelated) samples, as in the case of comparing two nonadjacent year estimates.

The first and second prevalence estimates may take the form of prevalence estimates from two survey years (e.g., 2023 and 2024, respectively), prevalence estimates from sets of combined survey data (e.g., 2021-2022 annual averages and 2023-2024 annual averages, respectively), or prevalence estimates for different populations of interest within a single survey year. Quick tests (where the correlation of 0 is assumed) are great tools for gaining a better understanding of published estimates; however, the results of these quick tests should be confirmed using NSDUH data and an appropriate programming language.

Under the null hypothesis, the test statistic  $t$  is a random variable that asymptotically follows a  $t$ -distribution for moderate to large sample sizes. Therefore, calculated values of  $t$ , along with the appropriate  $df$ , can be used to determine the corresponding probability level (i.e.,  $p$  value). **Whether testing for differences between years or from different populations within the same year, the covariance term in the formula for  $t$  (see Formula 7.1 earlier) will, in general, not be equal to 0.** An appropriate programming language, such as SUDAAN<sup>®</sup> Software for Statistical Analysis of Correlated Data (RTI International, 2020), can be

used to compute estimates of  $t$  along with the associated  $p$  values such that the covariance term is calculated by taking the sample design into account. A similar procedure and formula for  $t$  can be used for estimated totals, except when the alternative SE calculation was used (see Chapter 5). Whenever the SE for an estimated total was calculated outside of SUDAAN using the alternative SE estimation method, the corresponding test statistics also were computed outside of SUDAAN. SUDAAN along with auxiliary SAS<sup>®</sup> code (SAS Institute Inc., 2023), Stata<sup>®</sup> (StataCorp LP, 2017), SAS, and R (R Core Team, 2024) examples showing the computational methods for generating  $p$  values of estimates of  $t$  for means and totals can be found in Appendix A ([Exhibit A.2.4](#) through [Exhibit A.2.6](#), [Exhibit A.3.4](#) through [Exhibit A.3.6](#), [Exhibit A.4.4](#) through [Exhibit A.4.6](#), and [Exhibit A.5.4](#) through [Exhibit A.5.6](#)).

Under the null hypothesis, the test statistic with known variances asymptotically follows a standard normal ( $Z$ ) distribution. However, because the variances of the test statistic are estimated, its distribution is more accurately described by the  $t$ -distribution for finite sample sizes. A sufficiently large sample size is required for the asymptotic properties to take effect, and this is usually determined through the suppression criteria applied to the estimates (see Chapter 9). As the  $df$  approach infinity, the  $t$ -distribution approaches the  $Z$  distribution; **that is, because most of the statistical tests performed have 750  $df$  (see Chapter 6), the  $t$  tests performed produce approximately the same numerical results as if a  $Z$  test had been performed.**

Small differences in estimates between years can be statistically significant because of NSDUH's large sample sizes. These large sample sizes in each year reduce the size of the variances and increase the likelihood that the  $t$  test will yield a statistically significant difference. As stated previously, however, small differences between estimates that are not explained by sampling variability are not necessarily relevant from a policy perspective.

Caution is needed when interpreting changes across years in the estimated numbers of people with a characteristic of interest. Respondents with large analysis weights can greatly influence the estimated number in a given year when the number of people in the population with that characteristic is relatively small (e.g., past month heroin users). Large analysis weights for some respondents in a single year can result in the estimated numbers of people with a given characteristic showing an increase between year 1 and year 2 (i.e., the year that had the respondents with large analysis weights). The potential for these kinds of year-to-year variations in estimated numbers of people also underscores the importance of reviewing changes across a larger range of years when possible, especially for outcome measures corresponding to a relatively small proportion of the total population.

Caution also is needed when interpreting changes in estimated numbers of people. A change in the estimated number of people with a characteristic of interest could reflect a change in the size of the overall population. Therefore, changes in estimated numbers of people should be considered in conjunction with the corresponding estimated percentages because percentages will control for changes in both the number of people with the characteristic of interest and the total number of people in the population. If corresponding percentages are not available (e.g., for estimates of the number of past year initiates in national reports), caution should be taken

in interpreting increases over time, which may be explained by population increases rather than by true increases in the characteristic of interest.

**If SUDAAN or another programming language is not available to compute the standard  $t$  test, using published estimates can provide similar pairwise testing results.** When comparing prevalence estimates shown in the detailed tables with their SEs, independent  $t$  tests for the difference of proportions can be performed and usually will provide the same results as tests performed in SUDAAN (see Sections 7.1.1 and 7.1.2). However, where the  $p$  value is close to the predetermined level of significance, results may differ for two reasons: (1) the covariance term is included in the SUDAAN tests, whereas it is not included in independent  $t$  tests; and (2) the reduced number of significant digits shown in the published estimates may cause rounding errors in the independent  $t$  tests.

### 7.1.1 Example of Comparing Prevalence Estimates between Years

The following example tests estimates of lifetime cigarette use among young adults aged 18 to 25 between years 1 and 2. Assume a prevalence estimate of 45.9 percent for year 1 and 43.5 percent for year 2. The corresponding SEs are 0.55 percent for year 1 and 0.53 percent for year 2. Assuming that the source data are not available or the user does not have access to an appropriate programming language (i.e., SUDAAN), the  $t$  test Formula 7.2 can be used with the assumption that the correlation is 0. Note that

$$\text{var}(\hat{p}_i) = (\text{SE}(\hat{p}_i))^2, \quad (7.3)$$

$$t_{750} = \frac{45.9 - 43.5}{\sqrt{0.55^2 + 0.53^2 - 2(0)(0.55)(0.53)}} = 3.1422.$$

Using a  $t$  test to find the corresponding  $p$  value when  $t = 3.1422$  and  $df = 750$  results in  $p$  value = 0.0017. This is very close to a  $p$  value of 0.0028 calculated in SUDAAN. This example confirms that the difference between the year 1 estimate of 45.9 percent and the year 2 estimate of 43.5 percent is statistically significant at the 0.05 level. The calculated  $p$  value assuming the correlation is 0 is larger than the actual  $p$  value, which supports the earlier assertion that assuming the correlation is 0 results in a more conservative  $p$  value. Note, however, that this calculation could produce a smaller  $p$  value due to the use of rounded estimates from the table.

The following example uses Formula 7.2 with the unrounded estimates and the covariance from SUDAAN. The extra digits and the covariance change the  $t$ -score slightly, resulting in the  $p$  value of 0.0028. The  $t$  statistic from the below formula gives the same results as the test in SUDAAN:

$$t_{750} = \frac{45.86111372 - 43.53559846}{\sqrt{(0.5495147)^2 + (0.5271908)^2 - 2(-0.0401774591165442)(0.5495147)(0.5271908)}} = 2.9943.$$

In addition, the correlations between estimates in adjacent years are generally very small and positive, but a negative correlation is possible. Estimates with negative correlations will also be close to 0; thus, the differences in SUDAAN-calculated  $p$  values and  $p$  values calculated from estimates using the second  $t$  test formula provided earlier in this chapter (where the correlation is assumed to be 0) would still be minimal, such as the small differences shown in this section. **However, where the  $p$  value is close to the predetermined level of significance, results may differ.**

### 7.1.2 Example of Comparing Prevalence Estimates between Years in Excel

Using the same numbers presented in Section 7.1.1, this example uses Excel functions to produce the same  $p$  value produced in the previous example. The same assumption is made about the correlation (i.e., it is 0) and that the variance of the prevalence estimate is defined in Formula 7.3. The correlation of 0 results in the simplified formula shown below (additionally, the variances have been replaced by SEs squared).

$$t_{df} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(SE(\hat{p}_1))^2 + (SE(\hat{p}_2))^2}} \quad (7.4)$$

Excel can be used to set up a simple table (shown below) to compare prevalence estimates. Cells A2 through E2 are the known values input by the user. Cells F2 and G2 contain functions. This table could extend over several rows to aid in comparing many different pairs of prevalence estimates (i.e., data for columns A through E would have to be entered for each row, then the formulas in columns F and G could be copied for all rows).

	A	B	C	D	E	F	G
1	$p_1$	$p_2$	$SE(p_1)$	$SE(p_2)$	$df$	$t$	$p$ value
2	45.9	43.5	0.55	0.53	750	3.1422	0.0017

The standardized test statistic is found using the simplified Formula 7.4 for  $t_{df}$ .

	A	B	C	D	E	F	G
1	$p_1$	$p_2$	$SE(p_1)$	$SE(p_2)$	$df$	$t$	$p$ value
2	45.9	43.5	0.55	0.53	750	$=(A2-B2)/SQRT(C2^2+D2^2)$	0.0017

The Excel T.DIST.2T function then calculates the two-tailed Student's  $t$ -distribution, a continuous probability distribution.

	A	B	C	D	E	F	G
1	$p_1$	$p_2$	$SE(p_1)$	$SE(p_2)$	$df$	$t$	$p$ value
2	45.9	43.5	0.55	0.53	750	3.1422	$=T.DIST.2T(ABS(F2),E2)$

Alternatively, the Excel NORM.S.DIST function can be used to calculate the standard normal cumulative distribution function because the  $t$ -distribution approaches the  $Z$  distribution as the

$df$  approach infinity. Tests performed having 750  $df$  produce approximately the same numerical results as if a  $Z$  test had been performed. This function refers to the test statistic as  $Z$  and does not require the  $df$  input.

	A	B	C	D	E	F	G
1	$p_1$	$p_2$	$SE(p_1)$	$SE(p_2)$	$df$	$t$	$p$ value
2	45.9	43.5	0.55	0.53	750	3.1422	=2*(1-NORMSDIST(ABS(F2)))

The T.DIST.2T and NORM.S.DIST functions yield the same  $p$  value, 0.0017. Although not generated in all NSDUH publications, some publications do include sampling error in the form of 95 percent confidence intervals (CIs). In terms of testing for differences between prevalence estimates shown with 95 percent CIs, it is important to note that non-overlapping CIs indicate a significant difference at the 5 percent level, but overlapping intervals do not necessarily imply that there is no significant difference. For additional information, see Chapter 8, Schenker and Gentleman (2001), and Payton et al. (2003).

## 7.2 Comparing Prevalence Estimates in Categorical Subgroups

In addition to examining estimates between years, significance testing is used when comparing population subgroups within a single year.

### 7.2.1 Testing among Levels of Categorical Subgroups

For population subgroups defined by three or more levels of a categorical variable (e.g., region, with four levels), starting with a test of whether there is any distinction between levels is recommended, in order to first control the error level for multiple comparisons. For this purpose, log-linear chi-square tests of independence of the subgroup and the prevalence variables were conducted in SUDAAN. Although these tests are generally not published in the national report and tables, they can aid in report writing for NSDUH publications to verify statements implying significance, such as claiming that the prevalence for a measure of interest varies by age groups. See [Exhibit A.2.11](#) for example SUDAAN code, [Exhibit A.3.11](#) for Stata code, [Exhibit A.4.11](#) for SAS code, and [Exhibit A.5.11](#) for R code showing this type of testing. **If Shah's Wald  $F$  test (transformed from the standard Wald chi-square) indicated overall significant differences, the significance of each particular pairwise comparison of interest was tested using SUDAAN analytic procedures to properly account for the sample design (RTI International, 2020).** See [Exhibit A.2.12](#), [Exhibit A.3.12](#), [Exhibit A.4.12](#), and [Exhibit A.5.12](#) for pairwise testing examples.

This two-step procedure protected against inappropriate inferences being drawn due to the number of pairwise differences tested. For example, if the overall log-linear chi-square test of independence among age groups is statistically significant, then the following pairwise tests would be conducted between age group estimates: (1) 12 to 17 vs. 18 to 25, (2) 12 to 17 vs. 26 or older, and (3) 18 to 25 vs. 26 or older. For the pairwise comparison between estimates for adolescents aged 12 to 17 and young adults aged 18 to 25, the prevalence estimate and variance for adolescents become the first prevalence estimate and variance,  $\hat{p}_1$  and  $\text{var}(\hat{p}_1)$ ,

respectively. The prevalence estimate and variance for young adults become the second prevalence estimate and variance,  $\hat{p}_2$  and  $\text{var}(\hat{p}_2)$ , respectively. The covariance term is 0 because there is no overlap across age groups.

For consistency with the typical criteria for statistical testing in NSDUH, differences in age groups (and differences among most population subgroups) were deemed to be statistically significant if the  $p$  value was less than 0.05. One exception is testing among racial or ethnic groups. A more conservative level of 0.01 is sometimes used for these subgroups in special reports to increase the confidence that statistically significant differences reflect real differences in the population. The relatively large number of race/ethnicity subgroups being compared (seven) and their varying sample sizes might otherwise result in spurious determinations of significant differences that are due to sampling variability.

### 7.2.2 Testing among Levels of Categorical Subgroups: Age Adjustment

Age adjustment is a method used to standardize age distributions across subgroups. The age adjustment is used to estimate what the prevalence would be if the age distribution of each subgroup category were the same. These age-adjusted estimates are relative measures and not the actual prevalence estimate. Age-adjusted estimates were created by adjusting the weights (Klein & Schoenborn, 2001) so that the age distribution matched that of the 2000 U.S. standard population. This was in accordance with federal agency best practices (Surveillance, Epidemiology, and End Results Program, n.d.).

The testing procedures discussed in Section 7.2.1 are applied to the testing of age-adjusted estimates. Specifically, for subpopulations defined by three or more levels of a categorical variable (e.g., race/ethnicity), overall log-linear chi-square tests of independence of the subgroups are first performed. If the test result indicates statistical significance, then the significance of each particular pairwise comparison of interest is tested, as described in Section 7.2.1.

### 7.2.3 Significance Testing among Categorical Subdomains in Pooled Data from Multiple Years

Some NSDUH reports present estimates using pooled data from 2 or more survey years to improve the precision of estimates. These estimates represent annual averages across the number of years of data being pooled. For example, estimates in *NSDUH Population Statistics Reports: Binge Alcohol Use in the Past Month* (Center for Behavioral Health Statistics and Quality [CBHSQ], 2025g) are based on pooled 2022-2024 NSDUH data and reflect an annual average across those 3 years.

The testing procedures discussed for single-year data in Section 7.2.1 are applied to testing of estimates among subdomains using pooled data. Specifically, for subpopulations defined by three or more levels of a categorical variable (e.g., age group, race/ethnicity), overall log-linear chi-square tests of independence of the subgroups are first performed. If the test result indicates statistical significance, then the significance of each particular pairwise comparison of interest is tested, as described in Section 7.2.1.

### 7.2.4 Testing among a Subdomain and the Overall Population

Significance testing can also be conducted using SUDAAN to compare estimates for individual subgroups with the corresponding estimate among the overall population (e.g., all adults aged 18 or older vs. adults employed full time). Because this testing involves two overlapping domains, a stacked dataset that includes two records for each respondent in the overlap is needed for analysis. However, comparing estimates between a subgroup and the overall population increases the covariance in the denominator of the  $t$  test formula described in Section 7.1. Subtracting this covariance term from the sum of the variance terms for the individual estimates will decrease the size of the denominator and increase the size of the  $t$  statistic. For this reason, small differences between a subgroup and the overall population can be statistically significant. These significance test results are generally not published, but they can aid in report writing for NSDUH publications to verify statements implying significance, such as claiming that the prevalence for a measure of interest is higher or lower among a certain subpopulation when compared with the overall population. See [Exhibit A.2.10](#) for example SUDAAN code, [Exhibit A.3.10](#) for Stata code, [Exhibit A.4.10](#) for SAS code, and [Exhibit A.5.10](#) for R code showing this type of testing. This testing method can also be used for testing subgroups that partially overlap, such as those with private insurance coverage and those with Medicaid/Children's Health Insurance Program coverage.

## 7.3 Comparing Prevalence Estimates to Identify Linear Trends

In addition to comparing subpopulations for one year versus another year, it can also be useful to test the linear trend for all data points across all years of interest. *Linear trend testing can inform users about whether prevalence use has decreased, increased, or remained steady over the entire span of the years of interest or about changes in specific measures.*

Various methods can be used to test a linear trend. Linear trend testing is presented in the *Key Substance Use and Mental Health Indicators in the United States: Results from the 2024 National Survey on Drug Use and Health* report (CBHSQ, 2025f). These linear trend tests are implemented using the SUDAAN procedure DESCRIP with CONTRAST statements looking across years to evaluate change over time. This nonparametric method uses the  $t$  test, similar to the pairwise method used when testing means between years and between demographic levels. Instead of using PAIRWISE statements, type I errors (incorrectly producing significant differences) are controlled by using orthogonal polynomial coefficients in the CONTRAST statement. See Table 12.2 in the coding section of the *2024 National Survey on Drug Use and Health: Public Use File Data Users' Guide* (CBHSQ, 2025e), which contains the coefficients needed for linear testing across multiple years. Although pairwise testing gives detailed information for testing between 2 years, it does not perform as well for overall trend information and increases type I errors.

The DESCRIP procedure for linear testing is a good approximation to a model-based approach. The parametric model-based method is more flexible to measure a change in measurement over time when controlling for multiple covariates as needed. The modeling method can be used to estimate more specific measures, such as testing a year effect in a trend model that

adjusts for seasonal effects and redesign effects or comparing an estimate with an estimated forecast using data up to a specified year. The modeling method may yield a slightly different result from the DESCRIPT method under similar settings. The coding examples section in the 2024 Public Use File Data Users' Guide (CBHSQ, 2025e) shows examples of both methods for linear trend testing.

## 7.4 Impact of Rounding in Interpreting Testing Results

Prevalence estimates in the form of percentages are presented in the national reports and detailed tables and are rounded to the nearest 10th of a percent. Rounding of estimates needs to be taken into account when interpreting the results of tests for statistical significance because testing is done using unrounded estimates. **Testing between two rounded prevalence estimates can indicate significant or nonsignificant differences involving seemingly identical estimates.** The following example assumes that the SEs for each year are similar<sup>24</sup> and is provided to aid users in interpreting significance testing results:

The difference between the estimate in prior year A and the estimate in the current year is statistically significant. The difference between the estimate in prior year B and the estimate in the current year is not significant. However, the estimates for prior years A and B appear to be identical. For example, the estimate for lifetime heroin use among people aged 12 or older is 1.8 percent for years A, B, C, and D, but only the estimates from years A and C are significantly different from the current year estimate of 2.1 percent. Although the rounded estimates appear the same for all 4 years, the unrounded estimates are 1.7553 percent for year A, 1.8340 percent for year B, 1.8153 percent for year C, and 1.8488 percent for year D.

---

<sup>24</sup> The SEs are used in determining statistical significance for testing between years. See Section 7.1 for more information.

## 8. Confidence Intervals

In some National Survey on Drug Use and Health (NSDUH) publications, sampling error has been quantified using confidence intervals (CIs). CIs provide a scale to judge how close the sample statistic is likely to be to the true population parameter under repeated sampling. For example, although a 95 percent CI varies for each sample, it is expected to capture the true population parameter in 95 percent of samples. The CI includes one value above the estimate and one below and is determined by using the sampling distribution together with the standard error (SE). The SE measures the effects of sampling variability, and the sampling distribution provides an error multiplier for the particular confidence level (e.g., 95 percent). Samples with more variability will result in a larger spread in the CI, as will higher confidence levels. If using 95 percent CIs, it is important to note that although non-overlapping CIs indicate a significant difference at the 5 percent significance level, overlapping CIs do not necessarily imply the opposite: that there is no significant difference.

**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.

To illustrate the logit transformation method, let the proportion  $P_d$  represent the true proportion for a particular analysis domain  $d$ . Then the logit transformation of  $P_d$ , commonly referred to as the “log odds,” is defined as

$$L = \ln[P_d / (1 - P_d)], \quad (8.1)$$

where “ln” denotes the natural logarithm.

Letting  $\hat{p}_d$  be the estimate of the domain proportion, the log odds estimate becomes

$$\hat{L} = \ln[\hat{p}_d / (1 - \hat{p}_d)]. \quad (8.2)$$

The lower and upper confidence limits of  $L$  are formed with  $\sqrt{\text{var}(\hat{p}_d)} = \text{SE}(\hat{p}_d)$  as

$$A = \hat{L} - K \left[ \frac{\sqrt{\text{var}(\hat{p}_d)}}{\hat{p}_d(1 - \hat{p}_d)} \right], \quad (8.3)$$

$$B = \hat{L} + K \left[ \frac{\sqrt{\text{var}(\hat{p}_d)}}{\hat{p}_d(1 - \hat{p}_d)} \right], \quad (8.4)$$

where  $\text{var}(\hat{p}_d)$  is the variance estimate of  $\hat{p}_d$ , the quantity in brackets is a first-order Taylor series approximation of the SE of  $\hat{L}$ , and  $K$  is the critical value of the  $t$ -distribution associated with a specified level of confidence and degrees of freedom ( $df$ ). Section 8.1 shows an example producing 95 percent confidence limits for national estimates. See Chapter 6 for more details on what  $df$  should be used for various subpopulations in order to determine  $K$  appropriately.

Although the distribution of the logit-transformed estimate,  $\hat{L}$ , is asymptotically normal, the variance term in the CI is estimated, and a critical value from the  $t$ -distribution is therefore appropriate when calculating CIs. A sufficiently large sample size is required for the asymptotic properties to take effect, and this is usually determined through the suppression criteria applied to the estimates (see Chapter 9).

Applying the inverse logit transformation to  $A$  and  $B$  earlier yields a CI for  $\hat{p}_d$  as follows:

$$\hat{P}_{d,lower} = \frac{1}{1 + \exp(-A)}, \quad (8.5)$$

$$\hat{P}_{d,upper} = \frac{1}{1 + \exp(-B)}, \quad (8.6)$$

where “exp” denotes the inverse log transformation. The lower and upper CI endpoints for percentage estimates are obtained by multiplying the lower and upper endpoints of  $\hat{p}_d$  by 100.

The CI for the estimated domain total,  $\hat{Y}_d$ , as estimated by

$$\hat{Y}_d = \hat{N}_d \cdot \hat{p}_d, \quad (8.7)$$

is obtained by multiplying the lower and upper limits of the proportion CI by  $\hat{N}_d$ . For domain totals  $\hat{Y}_d$ , where  $\hat{N}_d$  (weighted population total) is nonfixed (see Chapter 5), the CI approximation assumes that the sampling variation in  $\hat{N}_d$  is negligible relative to the sampling variation in  $\hat{p}_d$ .

The following examples illustrate how to compute CIs using published prevalence estimates and SEs, and how to use published CIs to determine the SE of a prevalence estimate. **The CIs of totals cannot be computed using published estimates because this computation requires the weighted sum of the measures, which is most often not a published estimate.** Appendix A includes examples of how to compute the CIs of means and totals using various programming languages. See [Exhibit A.2.8](#) for example SUDAAN® Software for Statistical Analysis of Correlated Data code (RTI International, 2020), [Exhibit A.3.8](#) for Stata® code (StataCorp LP, 2017), [Exhibit A.4.8](#) for SAS® code (SAS Institute Inc., 2023), and [Exhibit A.5.8](#) for R code (R Core Team, 2024) on how to compute the CIs of the means and

totals. The Section 8.1 example computes CIs using the formulas shown earlier, the Section 8.2 example computes CIs using Excel, the Section 8.3 example shows how to use the CIs to compute SEs, and the Section 8.4 example shows how to use Excel to compute the SE from the CIs.

## 8.1 Example of Calculating Confidence Intervals Using Published Prevalence Estimates and Standard Errors

The following example illustrates how to determine the 95 percent CI when the prevalence estimate and SE are provided. Assuming a prevalence estimate of 3.5 percent and a corresponding SE of 0.11, this example uses Formulas 8.2-8.6 to determine the 95 percent CI.

The log odds estimate can be defined using Formula 8.2 as follows:

$$\hat{L} = \ln[0.035 / (1 - 0.035)] = -3.3168.$$

The upper and lower confidence limits of the log odds can then be defined using Formulas 8.3 and 8.4:

$$A = -3.3168 - 1.96 \left[ \frac{0.0011}{0.0338} \right] = -3.806, \text{ and}$$

$$B = -3.3168 + 1.96 \left[ \frac{0.0011}{0.0338} \right] = -3.2530.$$

Applying the inverse logit transformation yields the CIs'  $p$  using Formulas 8.5 and 8.6:

$$\hat{p}_{d,lower} = \frac{1}{1 + \exp(3.3806)} = 0.0329, \text{ and}$$

$$\hat{p}_{d,upper} = \frac{1}{1 + \exp(3.2530)} = 0.0372.$$

Rounding to two significant digits, the 95 percent CI is therefore 3.3 to 3.7 percent.

The same CI can be calculated using a programming language. Slight differences may occur due to rounding error caused by the reduced number of significant digits shown in published estimates. However, the results are usually close.

## 8.2 Example of Calculating Confidence Intervals in Excel Using Published Prevalence Estimates and Standard Errors

Using the same estimates presented in Section 8.1, this example uses Excel functions to produce the same CIs produced in the previous example. Excel can be used to set up a simple table (shown below) to produce the CI. Cells A2 through D2 are the known values input by the

user. Cells E2 and F2 contain functions. This table could extend over several rows to aid in producing many CIs (i.e., data for columns A through D would have to be entered for each row, then the formulas in columns E and F could be copied for all rows).

	A	B	C	D	E	F
1	$p_d$	$SE(p_d)$	$\alpha$	$df$	$p_{d,lower}$	$p_{d,upper}$
2	0.035	0.0011	0.05	750	0.0329	0.0372

The lower confidence limit is determined using the extended formula for  $\hat{p}_{d,lower}$ .

	A	B	C	D	E	F
1	$p_d$	$SE(p_d)$	$\alpha$	$df$	$p_{d,lower}$	$p_{d,upper}$
2	0.035	0.0011	0.05	750	=1/(1+EXP(-(LN(A2/(1-A2)) - T.INV.2T(C2,D2)*(B2/(A2*(1-A2))))))	0.0372

The upper limit is determined using the extended formula for  $\hat{p}_{d,upper}$ .

	A	B	C	D	E	F
1	$p_d$	$SE(p_d)$	$\alpha$	$df$	$p_{d,lower}$	$p_{d,upper}$
2	0.035	0.0011	0.05	750	0.0329	=1/(1+EXP(-(LN(A2/(1-A2)) + T.INV.2T(C2,D2)*(B2/(A2*(1-A2))))))

The 95 percent CI is 3.3 to 3.7 percent.

In the Excel formulas for  $\hat{p}_{d,lower}$  and  $\hat{p}_{d,upper}$ , the Excel function T.INV.2T calculates the inverse of the two-tailed Student's  $t$ -distribution, a continuous probability distribution. The function arguments are T.INV.2T (probability,  $df$ ), where probability is the probability (between 0 and 1) for which the user would want to evaluate the inverse of the two-tailed Student's  $t$ -distribution. This is also sometimes referred to as the alpha level. For 95 percent CIs, the alpha level is always 0.05. For 99 percent CIs, the alpha level is 0.01. The example uses 750  $df$  for a national estimate for NSDUH, but this could be adjusted for smaller areas of estimation.

### 8.3 Example of Calculating Standard Errors Using Published Confidence Intervals

This example illustrates how to determine the SE for an estimate when only the prevalence and 95 percent CI are provided. The example uses a prevalence estimate of 3.5 percent and a 95 percent CI of 3.3 to 3.7 percent and will show how to determine the SE for use in significance testing. This example uses formulas provided earlier to determine the SE for the prevalence estimate of 3.5 percent.

Formula 8.8 can be used to calculate  $A$  (lower CI for log odds estimate) by using the lower CI of the prevalence estimate ( $p$ ) calculated from Formula 8.5.

$$A = \ln \left( \frac{\hat{p}_{d,lower}}{1 - \hat{p}_{d,lower}} \right). \quad (8.8)$$

$$\ln\left(\frac{0.035}{1-0.035}\right) = -3.3168.$$

Using Formula 8.3 for  $A$  (lower limit of the log odds ratio), the following formula can be converted as shown in Formula 8.9 and used to get the SE:

$$SE(\hat{p}_d) = \frac{(A - \hat{L})[\hat{p}_d(1 - \hat{p}_d)]}{-K}. \quad (8.9)$$

Recall from the Section 8.1 example that  $L = -3.3168$ . Thus, the SE is computed as follows:

$$SE(\hat{p}_d) = \frac{(-3.3806 + 3.3168)[0.035(1 - 0.035)]}{-1.96} = 0.0011 \text{ or } 0.11 \text{ percent}$$

Using similar steps, the SE can be produced from the upper CI with the formulas below. The denominator is positive in the SE formula when using the upper CI.

$$B = \ln\left(\frac{\hat{p}_{d,upper}}{1 - \hat{p}_{d,upper}}\right), \quad (8.10)$$

$$SE(\hat{p}_d) = \frac{(B - \hat{L})[\hat{p}_d(1 - \hat{p}_d)]}{K}. \quad (8.11)$$

$$B = -3.2530, \text{ and } SE(\hat{p}_d) = 0.0011 \text{ or } 0.11 \text{ percent.}$$

**Sometimes, a reduced number of significant digits shown in published estimates may cause rounding errors when producing SEs from the lower or upper limits of the CIs.** This can result in SE estimates that differ from SEs computed directly from the data using a programming language. However, SEs calculated in these two ways will usually provide the same testing results, although results may differ when the  $p$  value is close to the chosen significance threshold.

## 8.4 Example of Calculating Standard Errors in Excel Using Published Confidence Intervals

Using the same estimates presented in Section 8.3, this example uses Excel functions to produce the same SEs from the previous example (i.e., the SUDAAN-generated SE). Excel can be used to set up a simple table (shown below) to produce the SE from the upper and lower limits of the CI. Cells A2 through D2 are the known values input by the user. Cell E2 contains the function to determine the SE. This table could extend over several rows to aid in producing many SEs (i.e., data for columns A through D would have to be entered for each row, then the formula in column E could be copied for all rows). Once the methods used in this example have

determined the SE from the CI, the methods shown in the Section 7.1.2 example can be used to perform independent  $t$  tests for differences of reported estimates in Excel.

Calculate the SE from the lower limit of the CI:

	A	B	C	D	E
1	$p_d$	$p_{d,lower}$	$\alpha$	$df$	$SE(p_d)$
2	0.035	0.0329	0.05	750	0.0011

$$SE(\hat{p}_d) = 0.0011 \text{ or } 0.11 \text{ percent.}$$

Similar to the Section 8.2 example, the Excel function T.INV.2T is used in the formula to determine the SE.

	A	B	C	D	E
1	$p_d$	$p_{d,lower}$	$\alpha$	$df$	$SE(p_d)$
2	0.035	0.0329	0.05	750	$=(((LN(B2/(1-B2)))-LN(A2/(1-A2)))*(A2*(1-A2)))/(-T.INV.2T(C2,D2)))$

Calculate the SE from the upper limit of the CI:

	A	B	C	D	E
1	$p_d$	$p_{d,upper}$	$\alpha$	$df$	$SE(p_d)$
2	0.035	0.0372	0.05	750	0.0011

$$SE(\hat{p}_d) = 0.0011 \text{ or } 0.11 \text{ percent.}$$

This also requires the use of the Excel function T.INV.2T (see details in Section 8.2).

	A	B	C	D	E
1	$p_d$	$p_{d,upper}$	$\alpha$	$df$	$SE(p_d)$
2	0.035	0.0372	0.05	750	$=(((LN(B2/(1-B2)))-LN(A2/(1-A2)))*(A2*(1-A2)))/(T.INV.2T(C2,D2)))$

**Remember that a reduced number of significant digits shown in published estimates may cause rounding errors when producing SEs.** This can result in SE estimates that differ when using the lower or upper limit from SEs computed directly from the data using a programming language. However, SEs calculated in these two ways will usually provide the same testing results, although results may differ when the  $p$  value is close to the chosen significance level.

## 9. Suppression of Estimates with Low Precision and Rules for Presentation of Estimates

Direct estimates from the National Survey on Drug Use and Health (NSDUH) that are designated as unreliable are not shown in reports or tables and are noted by asterisks (\*). The criteria used to define unreliability of direct estimates from NSDUH are based on

- the prevalence (for proportion estimates),
- relative standard error (RSE, defined as the ratio of the standard error [SE] divided by the estimate), and
- sample size.

These suppression criteria for various NSDUH estimates are summarized in [Table 9.1](#). The suppression criteria for the prevalence rates, shown at the top of [Table 9.1](#), were changed for products using 2024 NSDUH data, primarily for greater simplicity. The new suppression criteria for 2024 were also applied to estimates from 2021 to 2023 that were included in national reports and tables for the 2024 NSDUH. Consequently, some estimates from 2021 to 2023 that were suppressed in prior years may be published in reports and tables for the 2024 NSDUH. For documentation of the suppression rules used in earlier reports and tables, see Chapter 3 in the *2023 National Survey on Drug Use and Health (NSDUH): Methodological Summary and Definitions* (Center for Behavioral Health Statistics and Quality [CBHSQ], 2024).

**Table 9.1 Summary of 2024 NSDUH Suppression Rules**

Estimate	Suppress if any of the following conditions are met:
Estimated prevalence rate, $\hat{p}$ , with unweighted sample size of denominator, $n$	<p>(1) The estimated prevalence rate, <math>\hat{p}</math>, is 0 or 1, or</p> <p>(2) <math>\frac{SE(\hat{p})/\hat{p}}{-\ln(\hat{p})} &gt; .175</math> when <math>\hat{p} \leq .5</math>, or</p> <p><math>\frac{SE(\hat{p})/(1-\hat{p})}{-\ln(1-\hat{p})} &gt; .175</math> when <math>\hat{p} &gt; .5</math>, or</p> <p>(3) <math>n &lt; 50</math>.</p> <p>Rule (1) applies in the rare situations where estimated percentages are calculated as 100.0 percent because the numbers of respondents in the numerator and denominator are identical by chance. It does not apply in situations where reported percentages are exactly 100.0 percent because the numbers of respondents in the numerator and denominator have been forced to be identical.</p>
Estimated number (numerator of $\hat{p}$ )	The associated estimated prevalence rate, $\hat{p}$ , is suppressed.
Means other than percentages (e.g., mean age at first use), $\bar{x}$ , with sample size, $n$	<p>(1) <math>RSE(\bar{x}) &gt; .5</math>, or</p> <p>(2) <math>n &lt; 10</math>.</p>

RSE = relative standard error; SE = standard error.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2024.

Sample SAS code based on SAS and SUDAAN® (RTI International, 2020) output, Stata® code (StataCorp LP, 2017), SAS code (SAS Institute Inc., 2023), R code (R Core Team, 2024), and SPSS code (IBM Corp., 2017) demonstrating how to implement these rules can be found in Appendix A ([Exhibit A.2.3](#), [Exhibit A.3.3](#), [Exhibit A.4.3](#), [Exhibit A.5.3](#), and [Exhibit A.6.3](#), respectively).

## 9.1 Suppression Rules for Proportions

Under the current rules, direct estimates are suppressed under any of the following specific situations:

1. Prevalence estimates are suppressed if they are exactly 0 or 1 (or expressed in percentages as 0 or 100 percent), not counting the exception noted in [Table 9.1](#).
2. Proportion estimates ( $\hat{p}$ ), or rates, are suppressed if<sup>25</sup>

$$\frac{SE(\hat{p}) / \hat{p}}{-\ln(\hat{p})} > .175 \text{ when } \hat{p} \leq .5$$

or

$$\frac{SE(\hat{p}) / (1 - \hat{p})}{-\ln(1 - \hat{p})} > .175 \text{ when } \hat{p} > .5 .$$

3. Proportion estimates are also suppressed if the unweighted sample size of the denominator,  $n$ , is less than 50.

Under Rule 1, prevalence estimates of exactly 0 or 1 are suppressed because they will disclose information about all respondents in the domain.

Rule 2 suppresses prevalence estimates with large estimated SEs that indicate low precision. Rather than a simple RSE, the suppression rule for proportions based on  $RSE[-\ln(\hat{p})]$  is used because it ensures that the strictness of the rule is similar for both large and small values of  $\hat{p}$ . Additionally, the separate formulas for  $\hat{p} \leq .5$  and  $\hat{p} > .5$  produce a symmetric suppression rule; that is, if  $\hat{p}$  is suppressed,  $1 - \hat{p}$  also will be suppressed. In contrast, a commonly used rule that suppresses estimates when  $RSE(\hat{p}) > .5$  is stringent when  $\hat{p}$  is small (e.g., 0.01) but much less so when  $\hat{p}$  is large (e.g., 0.49). In addition, a rule based on only  $RSE(\hat{p}) > .5$  is

<sup>25</sup> The computational formula in Rule 2 was derived from  $RSE[-\ln(\hat{p})]$ :  $RSE[-\ln(\hat{p})] \equiv SE[-\ln(\hat{p})] / [-\ln(\hat{p})]$  for  $\hat{p} \leq 0.5$ . The Taylor series linearization of the numerator  $SE[-\ln(\hat{p})]$  is  $SE[-\ln(\hat{p})] = \sqrt{\text{var}[-\ln(\hat{p})]}$ , which approximately equals  $\sqrt{(-1/\hat{p})^2 \text{var}(\hat{p})}$  by Taylor series linearization, which in turn equals  $SE(\hat{p})/\hat{p}$ . The same principles apply for the computational formula when  $\hat{p} > 0.5$ , except that  $\hat{p}$  is replaced with  $1 - \hat{p}$ .

asymmetric in the sense that suppression occurs only in terms of  $\hat{p}$ ; that is, there is no complementary rule for  $(1 - \hat{p})$ . This would mean that changing the way that the question is formulated would result in different suppression decisions, even when the information is the same.

Under Rule 3, estimates are also suppressed if the sample size is less than 50 in order to protect against the instability of the estimated SE for small sample sizes. Unstable estimated SEs cause Rule 2 to be unreliable.

## 9.2 Suppression Rule for Estimated Totals, Means, and Sample Sizes

Estimates of totals (i.e., estimated numbers of people) were suppressed if the corresponding prevalence rates were suppressed. Because of this rule, data users may encounter some unexpected results after applying the suppression rules. For instance, equivalent estimates of totals corresponding to different estimated percentages are suppressed differently.

Consider a situation where estimates of the misuse of prescription drugs in the past year are presented among the population aged 12 or older and among people who used prescription drugs for any reason in the past year. Because the associated percentages have different denominators,  $\hat{p}$  may not be suppressed for the population aged 12 or older but could be suppressed for the percentage among past year users. In this situation, the estimated total would be displayed for the population aged 12 or older. However, the same estimated total that is associated with the suppressed percentage among past year users would be suppressed. For example, Table 1.22 in the *Results from the 2022 National Survey on Drug Use and Health: Detailed Tables* (CBHSQ, 2023b) shows among the total population that an estimated 203,000 adolescents aged 12 to 17 in 2021 misused benzodiazepines in the past year. That estimated number was shown as being suppressed for misuse among people who used benzodiazepines for any reason in the past year because the corresponding percentage was suppressed for benzodiazepine misuse among people who used benzodiazepines for any reason.

Estimates of means that are not bounded between 0 and 1 (e.g., mean of age at first use, mean number of days of use in the past 30 days or the past 12 months) were suppressed if the RSEs of the estimates were larger than .5 or if the nominal sample size was smaller than 10 respondents. This rule was based on an empirical examination of the estimates of mean age of first use and their SEs for various empirical sample sizes. Although arbitrary, a sample size of 10 appeared to provide sufficient precision and still allow reporting by age at first use for many substances.

### 9.3 Rounding Rules and Format for Presentation of Certain Unsuppressed Estimates

In addition to the suppression rules, the NSDUH national tables and reports present rounded estimates and apply specific formats to estimates in order to provide additional confidentiality protection. [Table 9.2](#) describes these rounding and presentation rules.

**Table 9.2 Rounding Rules and Format for 2024 NSDUH Tables and Reports**

Estimate Type	Rounding and Presentation Format
Estimated Totals and Standard Errors of Totals	<ul style="list-style-type: none"> <li>Total estimates are rounded to the nearest thousand people.</li> <li>Estimated numbers less than 500 people are shown as "&lt;1" in tables to indicate that the number rounds to less than 1,000 people.</li> <li>Estimated numbers greater than or equal to 500 but less than 1,000 people are shown as "1" in tables because they round to 1,000 people.</li> </ul>
Estimated Percentages	<ul style="list-style-type: none"> <li>Estimated percentages are rounded to the nearest tenth of a percent.</li> <li>Estimated percentages less than 0.05 are shown as "&lt;0.1" in tables.</li> <li>Estimated percentages greater than or equal to 0.05 but less than 0.1 are shown as "0.1" in tables because they round to 0.1 percent.</li> </ul>
Standard Errors of Percentages	<ul style="list-style-type: none"> <li>Estimated standard errors of percentages are rounded to the nearest hundredth of a percent.</li> <li>Estimated standard errors of percentages less than 0.005 are shown as "&lt;0.01" in tables.</li> <li>Estimated standard errors of percentages greater than or equal to 0.005 but less than 0.01 are shown as "0.01" in tables because they round to 0.01.</li> </ul>
Unweighted Sample Sizes <sup>1</sup>	<ul style="list-style-type: none"> <li>Unweighted sample sizes are rounded to the nearest 10 people.</li> <li>Unweighted sample sizes less than 100 people are shown as "&lt;100" in tables.</li> </ul>
PValues	<ul style="list-style-type: none"> <li>P values for the test of differences between estimates are rounded to the nearest ten thousandth.</li> </ul>

<sup>1</sup> This suppression rule was implemented for confidentiality protection. The other suppression rules did not apply because no mean is associated with the sample size and population estimates; thus, most of the components of the suppression criteria are not applicable. Also, because no behavior associated with the numbers is displayed, there is no risk of behavior disclosure.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2024.

## References

- Center for Behavioral Health Statistics and Quality. (2022). *2021 National Survey on Drug Use and Health (NSDUH): Methodological summary and definitions*.  
<https://www.samhsa.gov/data/report/2021-methodological-summary-and-definitions>
- Center for Behavioral Health Statistics and Quality. (2023a). *2022 National Survey on Drug Use and Health (NSDUH): Methodological summary and definitions*.  
<https://www.samhsa.gov/data/report/2022-methodological-summary-and-definitions>
- Center for Behavioral Health Statistics and Quality. (2023b). *Results from the 2022 National Survey on Drug Use and Health (NSDUH): Detailed tables*.  
<https://www.samhsa.gov/data/report/2022-nsduh-detailed-tables>
- Center for Behavioral Health Statistics and Quality. (2024). *2023 National Survey on Drug Use and Health (NSDUH): Methodological summary and definitions*.  
<https://www.samhsa.gov/data/report/2023-methodological-summary-and-definitions>
- Center for Behavioral Health Statistics and Quality. (2025a). *2024 Companion infographic report: Results from the 2021 to 2024 National Surveys on Drug Use and Health: (HHS Publication No. PEP25-07-006)*. <https://www.samhsa.gov/data/report/2024-nsduh-companion-infographic-report>
- Center for Behavioral Health Statistics and Quality. (2025b). *2024 National Survey on Drug Use and Health (NSDUH) methodological resource book, Section 2: Sample design and experience report*. <https://www.samhsa.gov/data/report/nsduh-2024-methodological-resource-book-mrb>
- Center for Behavioral Health Statistics and Quality. (2025c). *2024 National Survey on Drug Use and Health (NSDUH): Methodological summary and definitions*.  
<https://www.samhsa.gov/data/report/2024-methodological-summary-and-definitions>
- Center for Behavioral Health Statistics and Quality. (2025d). *2024 National Survey on Drug Use and Health public use file codebook*. Substance Abuse and Mental Health Services Administration.
- Center for Behavioral Health Statistics and Quality. (2025e). *2024 National Survey on Drug Use and Health: Public use file data users' guide*. <https://www.samhsa.gov/data/data-we-collect/nsduh/datafiles>
- Center for Behavioral Health Statistics and Quality. (2025f). *Key substance use and mental health indicators in the United States: Results from the 2024 National Survey on Drug Use and Health (HHS Publication No. PEP25-07-007, NSDUH Series H-60)*.  
<https://www.samhsa.gov/data/report/2024-nsduh-annual-national-report>
- Center for Behavior Health Statistics and Quality. (2025g). *NSDUH population statistics reports: Binge alcohol use in the past month (Vol. 1, No. 8)*.  
<https://www.samhsa.gov/data/report/nsduh-2022-2024-binge-alcohol-use-past-month>

Center for Behavioral Health Statistics and Quality. (2025h). *Results from the 2024 National Survey on Drug Use and Health (NSDUH): Detailed tables*.

<https://www.samhsa.gov/data/report/2024-nsduh-detailed-tables>

Center for Behavioral Health Statistics and Quality. (forthcoming a). *2024 National Survey on Drug Use and Health (NSDUH) methodological resource book, Section 10: Editing and imputation report*. Substance Abuse and Mental Health Services Administration.

Center for Behavioral Health Statistics and Quality. (forthcoming b). *2024 National Survey on Drug Use and Health (NSDUH) methodological resource book, Section 11: Person-level sampling weight calibration*. Substance Abuse and Mental Health Services Administration.

Formplus Blog. (2023, January 12). Population of interest – Definition, determination, comparisons. *Formplus*. <https://www.formpl.us/blog/population-of-interest>

IBM Corp. (2017). *IBM SPSS statistics for windows*. <https://hadoop.apache.org>

Klein, R. J., & Schoenborn, C. A. (2001). Age adjustment using the 2000 projected U.S. population. *Healthy People Statistical Notes, 20*. National Center for Health Statistics. <https://www.cdc.gov/nchs/data/statnt/statnt20.pdf>

Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science, 3*, 34. <https://doi.org/10.1673/031.003.3401>

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

RTI International. (2020). *SUDAAN® language manual, release 11.0.4*.

SAS Institute Inc. (2023). *SAS/STAT software: Release 15.3*.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician, 55*(3), 182-186. <https://doi.org/10.1198/000313001317097960>

StataCorp LP. (2017). *Stata statistical software: Release 14*.

Surveillance, Epidemiology, and End Results Program. (n.d.). *Use of the 2000 U.S. standard population for age-adjustment*. National Cancer Institute. <https://seer.cancer.gov/stdpopulations/2000stdpop-use.html>

## 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. Elizabeth Crane served as contracting officer representative, Carlos Graham served as assistant contracting officer representative, and David Hunter served as RTI project director.

Contributors at RTI included Kristen Gullede Brown and Teresa R. Davis.

## Appendix A: Documentation for Conducting Various Statistical Procedures: SUDAAN®, Stata®, SAS®, R, and SPSS Examples

This appendix provides guidance concerning various options that should be specified in SUDAAN® Software for Statistical Analysis of Correlated Data (RTI International, 2020), Stata® (StataCorp LP, 2017), SAS® (SAS Institute Inc., 2023), R (R Core Team, 2024), and SPSS (IBM Corp, 2017) to correctly analyze the National Survey on Drug Use and Health (NSDUH) data. Example SUDAAN, Stata, SAS, and R code is provided to illustrate how the information in this report is applied to generate estimates (means, totals, and percentages, along with standard errors [SEs]), implement the suppression rule, perform statistical tests of differences, handle missing data, calculate confidence intervals (CIs), test between overlapping domains, test independence of two variables, and perform pairwise tests. Example SPSS code is provided to illustrate how the information in this report is applied to generate estimates (means, totals, and percentages, along with SEs) and implement the suppression rule.

Specifically, Section A.1 provides overall guidance on how to use various programming options for NSDUH within the five software packages. In the remaining sections of this appendix, these options are then used within specific examples of how to produce estimates using the statistical procedures documented within this report; there are separate sections for each programming language. The examples are created using variable names found on the restricted-use data file; some variable names may be different from those found on the public use data file.<sup>4</sup>

All estimates in the *Results from the 2024 National Survey on Drug Use and Health: Detailed Tables* (Center for Behavioral Health Statistics and Quality [CBHSQ], 2025h) and the *Key Substance Use and Mental Health Indicators in the United States: Results from the 2024 National Survey on Drug Use and Health* report (CBHSQ, 2025f) are produced using survey analysis procedures in SUDAAN and auxiliary SAS code. However, the Stata, SAS, R, and SPSS survey analysis code creates equivalent results. Results may vary slightly across programming languages because of differences in precision or the default degrees of freedom.

The appendix section number for each programming language, the exhibit number for each example within that section, a description of the example, and a reference to the report chapter that addresses the example are provided in [Table A.1](#).

**Table A.1 Summary of SUDAAN, Stata, SAS, R, and SPSS Exhibits**

Section A.2: SUDAAN/SAS Exhibits	Section A.3: Stata Exhibits	Section A.4: SAS Exhibits	Section A.5: R Exhibits	Section A.6: SPSS Exhibits	Description	Report Chapter
<a href="#">Exhibit A.2.1</a>	<a href="#">Exhibit A.3.1</a>	<a href="#">Exhibit A.4.1</a>	<a href="#">Exhibit A.5.1</a>	<a href="#">Exhibit A.6.1</a>	Produces estimates (including means, totals, and the respective SEs) using single- or combined-year (pooled) data.	Chapters 3, 5, and 6
<a href="#">Exhibit A.2.2</a>	<a href="#">Exhibit A.3.2</a>	<a href="#">Exhibit A.4.2</a>	<a href="#">Exhibit A.5.2</a>	<a href="#">Exhibit A.6.2</a>	Calculates the SE of the total for fixed domains using the alternative SE estimation method using the estimates produced in <a href="#">Exhibit A.2.1</a> , <a href="#">Exhibit A.3.1</a> , <a href="#">Exhibit A.4.1</a> , <a href="#">Exhibit A.5.1</a> , and <a href="#">Exhibit A.6.1</a> .	Chapter 5
<a href="#">Exhibit A.2.3</a>	<a href="#">Exhibit A.3.3</a>	<a href="#">Exhibit A.4.3</a>	<a href="#">Exhibit A.5.3</a>	<a href="#">Exhibit A.6.3</a>	Creates suppression indicators for each estimate (i.e., suppression rule).	Chapter 9
<a href="#">Exhibit A.2.4</a>	<a href="#">Exhibit A.3.4</a>	<a href="#">Exhibit A.4.4</a>	<a href="#">Exhibit A.5.4</a>	--	Performs statistical tests of differences between means.	Chapter 7
<a href="#">Exhibit A.2.5</a>	<a href="#">Exhibit A.3.5</a>	<a href="#">Exhibit A.4.5</a>	<a href="#">Exhibit A.5.5</a>	--	Calculates the $p$ value for the test of differences between totals of nonfixed domains (using estimates produced in <a href="#">Exhibit A.2.4</a> , <a href="#">Exhibit A.3.4</a> , <a href="#">Exhibit A.4.4</a> , and <a href="#">Exhibit A.5.4</a> ).	Chapter 7
<a href="#">Exhibit A.2.6</a>	<a href="#">Exhibit A.3.6</a>	<a href="#">Exhibit A.4.6</a>	<a href="#">Exhibit A.5.6</a>	--	Calculates the $p$ value for the test of differences between fixed domains by producing the covariance matrix, pulling the relevant covariance components, and calculating the variances.	Chapter 7
<a href="#">Exhibit A.2.7</a>	<a href="#">Exhibit A.3.7</a>	<a href="#">Exhibit A.4.7</a>	<a href="#">Exhibit A.5.7</a>	--	Produces estimates where the variable of interest has missing values.	Chapter 4
<a href="#">Exhibit A.2.8</a>	<a href="#">Exhibit A.3.8</a>	<a href="#">Exhibit A.4.8</a>	<a href="#">Exhibit A.5.8</a>	--	Calculates confidence interval using estimates produced in <a href="#">Exhibit A.2.1</a> , <a href="#">Exhibit A.3.1</a> , <a href="#">Exhibit A.4.1</a> , and <a href="#">Exhibit A.5.1</a> .	Chapter 8
<a href="#">Exhibit A.2.9</a>	<a href="#">Exhibit A.3.9</a>	<a href="#">Exhibit A.4.9</a>	<a href="#">Exhibit A.5.9</a>	--	Calculates percentages and the associated SEs for a categorical measure of interest.	Chapters 3 and 5
<a href="#">Exhibit A.2.10</a>	<a href="#">Exhibit A.3.10</a>	<a href="#">Exhibit A.4.10</a>	<a href="#">Exhibit A.5.10</a>	--	Performs statistical tests of differences between two groups when the two groups overlap.	Chapter 7
<a href="#">Exhibit A.2.11</a>	<a href="#">Exhibit A.3.11</a>	<a href="#">Exhibit A.4.11</a>	<a href="#">Exhibit A.5.11</a>	--	Performs tests of the independence of the prevalence variable and subgroup variable.	Chapter 7
<a href="#">Exhibit A.2.12</a>	<a href="#">Exhibit A.3.12</a>	<a href="#">Exhibit A.4.12</a>	<a href="#">Exhibit A.5.12</a>	--	Performs pairwise tests for each level of the subgroup variable.	Chapter 7
<a href="#">Exhibit A.2.13</a>	<a href="#">Exhibit A.3.13</a>	<a href="#">Exhibit A.4.13</a>	<a href="#">Exhibit A.5.13</a>	--	Performs linear test of significance across years using test statements	Chapter 7

-- Not available.  
SE = standard error.

## A.1 Guide for Defining Options for Analyzing NSDUH Data

The programming languages shown in [Table A.1](#) can be used to generate weighted estimates, unweighted sample sizes, means, totals, SEs of means and totals, and  $p$  values for testing of the means and totals. The options described in Section A.1 are used within the SUDAAN, Stata, SAS, R, and SPSS examples to correctly produce estimates using NSDUH data. Because the SPSS examples only include code to generate estimates (means, totals, and percentages, along with SEs) and implement the suppression rule, some of the options may not apply, and if they do not apply, the SPSS information is not provided.

Before running the SUDAAN procedures, the input dataset must be sorted by the nesting variables (e.g., VESTR 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. Stata, SAS, R, and SPSS commands can be run without the data being sorted.

### A.1.1 NSDUH Sample Design

The NSDUH sample design is complex, and it is based on multistage clustered sample designs where the first-stage (primary) sampling units are drawn with replacement. See Section 2.1 for more information on the sample design.

It is important to account for the sample design when analyzing NSDUH data to get proper estimates of variance. The programming languages shown in [Table A.1](#) vary with respect to how they do this. In SUDAAN, NSDUH estimates are calculated based on the Taylor series linearization method that is unbiased for linear statistics by specifying DESIGN=WR (meaning “with replacement”). In SAS, the SURVEYMEANS procedure can specify the VARMETHOD=TAYLOR option or not specify the VARMETHOD= option in which the Taylor linearized variance estimation is the default. In Stata, the Taylor linearized variance estimation is also the default, but other options can be specified.

The R sample codes in this appendix mainly use the *survey* package. Users must install it before trying to run the sample codes. In addition to the *survey* package, three more R packages are used. [Exhibit A.5.1](#) (the first R example) starts with installing all necessary packages. The version information is as follows: The sample code blocks were tested in R version 4.4-8, with survey 4.4-2, haven 2.5.5, dplyr 1.1.4, and multcomp 1.4-29.

The R *survey* package allows the use of both Taylor series linearization and replication weighting for variance estimation. The sample codes in this appendix use only the former one. The object, *svydesign*, specifies a complex survey design. The *svydesign* object combines data and all the survey design information needed to analyze the data.

The SPSS sample codes in this appendix mainly use the CSDESCRIPTIVES command. This command requires the PLAN subcommand that specifies the name of an XML design file detailing various specifications of the survey design. This XML file can be created using the CSPLAN command as demonstrated in [Exhibit A.6.1](#). The remaining SPSS exhibits assume the XML design file has already been created.

### A.1.2 Nesting Variables

In the examples below, the NSDUH nesting variables (VESTR and VEREP) are used to capture explicit stratification and to identify clustering within the NSDUH data, which are needed to compute the variance estimates correctly. See Section 2.1 for more details on stratification. Two replicates per year were defined within each variance stratum (VESTR). Each variance replicate (VEREP) consists of four segments, one for each quarter of data collection. One replicate consists of those segments that are “phasing out” or will not be used in the next survey year. The other replicate consists of those segments that are “phasing in” or will be fielded again the following year, except in 2024, when the “phasing in” segments will not be used again. A segment stays in the same VEREP for the 2 years it is in the sample. This simplifies computing SEs for estimates based on combined data from adjacent survey years.

In SUDAAN, users must use the NEST statement within one of the appropriate SUDAAN procedures. In the NEST statement, the variable for the variance stratum should be listed first, followed by the primary sampling unit variable; that is, the VESTR variable should always be listed first, followed by the VEREP variable. In Stata, the nesting variables are specified in the svyset command. In SAS, users must use the STRATA and CLUSTER statements within one of the appropriate SAS procedures. VESTR should be listed in the STRATA statement, and VEREP should be listed in the CLUSTER statement. Unlike the svyset command in Stata where it needs to be called only once, the NEST statement in SUDAAN and the STRATA and CLUSTER statements in SAS will need to be used each time a user calls one of the appropriate SUDAAN or SAS procedures, respectively. Similar to Stata, once the survey design object is created with svydesign in R, the object is used subsequently with no need for repeating. In SPSS, once the design file is created, it can be used subsequently with no need for creating another design file.

### A.1.3 Degrees of Freedom

As described in Chapter 6, the degrees of freedom (DDF in SUDAAN and dof in Stata) are 750 for the 2024 national estimates: 144 in California; 120 each in Florida, New York, and Texas; 96 each in Illinois, Michigan, Ohio, and Pennsylvania; 60 each in Georgia, New Jersey, North Carolina, and Virginia; and 48 each in the remaining 38 states and the District of Columbia. Due to potential overlap of variance strata, analyses for a group of states may need degrees of freedom less than the sum of the degrees of freedom for each individual state. The specific number of degrees of freedom can be computed by counting the unique values of VESTR for the particular geographic area of interest. The technique of counting the number of unique values of VESTR is also applicable for any analyses combining survey data across years. For more information on degrees of freedom for various types of analyses, see [Table 6.3](#) in Chapter 6.

To specify the degrees of freedom in SUDAAN, the DDF = option on the procedure statement is used. This option should be used each time one of the appropriate SUDAAN procedures is called to ensure correct calculations. In Stata, the degrees of freedom are specified as a design option in the svyset command; that is, “dof(750).” If switching from national estimates to state estimates, the svyset command would need to be rerun with the updated degrees of freedom. In SAS, degrees of freedom can be specified by using the DF = option in the MODEL statement.

For the Taylor series method, the default degrees of freedom equal the number of clusters minus the number of strata (in this case, 1,500 – 750). In R, if DF are not provided, the DF are estimated from a model. A function, `degf`, is used to extract DF from a survey design. See an example of using `degf` in the R sample codes.

#### A.1.4 Design Effect

When generating estimates based on survey data, it is important to account for the survey design. Accounting for the survey design elements such as stratification (or blocking), clustering, and unequal weighting allows for correct variance calculations. The option DEFT4 within SUDAAN provides the correct measure of variance inflation. Requesting `deff srssubpop` in Stata, `DEFF="REPLACE"` in R, or `DEFF` in SPSS gives the same result as using DEFT4 in SUDAAN. The design effect cannot be output directly from the SURVEYMEANS procedure in SAS. In the SAS exhibits, the UNIVARIATE procedure with the `VARDEFF=WGT` option is used to correctly calculate the variance under simple random sampling.

#### A.1.5 Standard Errors

As discussed in Chapter 5, sampling error is the error caused by using estimates based on a sample instead of true values (parameters) based on a census. SEs are commonly used to measure the uncertainty of the estimates. The SE for the mean (or proportion) comes directly out of SUDAAN, SAS, R, and SPSS in the output variables SEMEAN ([Exhibit A.2.1](#)), STDERR ([Exhibit A.4.1](#)), SE (using SVYBY in [Exhibit A.5.1](#)), and StandardError where `Var1='Mean'` ([Exhibit A.6.1](#)), respectively, and the SEMEAN is calculated in Stata by taking the square root of the variance ([Exhibit A.3.1](#)).

However, to compute the SEs of the totals, NSDUH implements different methods depending on whether the specified domain (i.e., sex in this example) is fixed or nonfixed. For the 2024 NSDUH, [Table 5.1](#) contains a list of fixed domains. If a domain is nonfixed (e.g., not forced to match the U.S. Census Bureau or American Community Survey [ACS] population estimates), then the SE of the total comes directly out of SUDAAN, SAS, R, and SPSS in the output variables SETOTAL, STDDEV, SE (using svymean in [Exhibit A.5.1](#)), and StandardError where `Var1='sum'`, respectively. If the domain is fixed (e.g., forced to match the U.S. Census Bureau population estimates), then the SE of the total is calculated using an alternative SE estimation method; that is, SETOTAL (SE of fixed domain) = WSUM (weighted sample size) × SEMEAN (SE for the mean/proportion), as seen in [Exhibit A.2.2](#), [Exhibit A.3.2](#), [Exhibit A.4.2](#), [Exhibit A.5.2](#), and [Exhibit A.6.2](#).

#### A.1.6 Suppression Rule

As described in Chapter 9, each published NSDUH estimate goes through a suppression rule to detect whether the estimate is unreliable because of an unacceptably large sampling error. Starting with the 2024 NSDUH, the suppression rule was simplified. The suppression rules as they apply to different types of estimates are shown in [Table 9.1](#). The examples in [Exhibit A.2.3](#) (SAS code based on SUDAAN output), [Exhibit A.3.3](#) (Stata code), [Exhibit A.4.3](#) (SAS code), [Exhibit A.5.3](#) (R code), and [Exhibit A.6.3](#) (SPSS code) show the prevalence estimate rule and

the rule for means not bounded by 0 and 1 (i.e., averages). The average suppression rule is commented out for these examples, but it would replace the prevalence estimate suppression rule if averages were shown in the examples in place of means bounded by 0 and 1.

For tables that display totals along with multiple means from differing populations (e.g., initiation tables in Section 4 of the 2024 Detailed Tables [CBHSQ, 2025h]), suppression is not as straightforward as coding the rule in the SAS/SUDAAN, Stata, R, or SPSS programs. As discussed in Chapter 9, there may be instances where some means are suppressed, and others are not suppressed. In that instance, suppression of the total estimate is based on the level of suppression present across all corresponding mean estimates. If all mean estimates associated with a total estimate are suppressed, the total estimate should also be suppressed. If at least one mean estimate is not suppressed, the total estimate is also not suppressed.

### A.1.7 Statistical Tests of Differences between Years

As described in Chapter 7, significance tests can be conducted on differences of prevalence estimates between years and among subdomains within years. For the 2024 NSDUH, statistical testing was done between 2023 and 2024. Sample code is provided for reference. If year-to-year tests of differences are computed and the estimate for either year is suppressed, then the resulting  $p$  value shown in the P tables within the detailed tables would also be suppressed. This is the rule used when creating the detailed tables; however, this code does not show this rule being implemented.

For the SUDAAN example ([Exhibit A.2.4](#)), testing of differences requires a separate PROC DESCRIPT run from the initial DESCRIPT run that produces the corresponding yearly estimates. Tests of differences can be generated using DESCRIPT's CONTRAST, PAIRWISE, or DIFFVAR statements. The SUDAAN example ([Exhibit A.2.4](#)) uses the DIFFVAR statement to test for differences between a pair of years of past month alcohol use estimates for all people aged 12 or older (IRSEX=0), all males (IRSEX=1), and all females (IRSEX=2). It also includes an example of using multiple DIFFVAR statements to test for differences between any two years. Similarly, for the Stata example ([Exhibit A.3.4](#)), a separate svy: mean command is needed. The SAS procedure SURVEYREG is used to compute the test of differences. R uses SVYTTTEST from the *survey* package. No examples using SPSS to compute statistical testing are currently provided.

Similar to computing the SEs of the totals, calculating  $p$  values for tests of differences of totals differs depending on whether an estimate is considered to be from a fixed domain or a nonfixed domain. Both ways are described as follows with accompanying example code: [Exhibit A.2.4](#) and [Exhibit A.2.5](#) show example code for nonfixed domains using SUDAAN and auxiliary SAS, [Exhibit A.3.4](#) and [Exhibit A.3.5](#) show the same example using Stata, [Exhibit A.4.4](#) and [Exhibit A.4.5](#) show the example using SAS, and [Exhibit A.5.4](#) and [Exhibit A.5.5](#) show the example using R. [Exhibit A.2.6](#) shows example code for fixed domains using SUDAAN and auxiliary SAS. [Exhibit A.3.6](#) shows example code for fixed domains using Stata. [Exhibit A.4.6](#) shows the example using SAS, and [Exhibit A.5.6](#) shows the example using R. See Section A.1.11 for information about testing subdomains with years.

### A.1.8 Recoding and Missing Values

Missing data in NSDUH variables should often be excluded from the analysis. In the examples in [Exhibit A.2.7](#) (using auxiliary SAS and SUDAAN), [Exhibit A.3.7](#) (using Stata), [Exhibit A.4.7](#) (using SAS), and [Exhibit A.5.7](#) (using R), the mean age of first use of marijuana will be calculated in two ways in each exhibit. Respondents who have never used marijuana are assigned IRMJAGE=991, and if this level is included in the analysis, then the mean age calculated will be too high. Thus, two methods are shown on how to omit this level in calculating mean age of first use of marijuana. The first example sets the level of 991 to system missing, and that is automatically excluded from the analysis. The second example subsets the data to remove the level of 991 from the analysis.

### A.1.9 Confidence Intervals

As discussed in Chapter 8, CIs provide a scale to judge how close the sample statistic is likely to be to the true population parameter under repeated sampling and can be calculated using means (MEAN) and SEs (SEMEAN) from PROC DESCRIPT in SUDAAN, svy: mean in Stata, the SURVEYMEANS procedure in SAS, and svymean in R. After the means and SEs are obtained ([Exhibit A.2.2](#), [Exhibit A.3.2](#), [Exhibit A.4.2](#), and [Exhibit A.5.2](#)), the code in [Exhibit A.2.8](#), [Exhibit A.3.8](#), [Exhibit A.4.8](#), and [Exhibit A.5.8](#) can be used to create the 95 percent CIs for means and totals.

### A.1.10 Calculating Percentages for Categories

[Exhibit A.2.9](#), [Exhibit A.3.9](#), [Exhibit A.4.9](#), and [Exhibit A.5.9](#) demonstrate how to compute estimates corresponding to levels of a categorical variable. This example uses the number of days used marijuana in the past month among past month marijuana users. The variable that will be analyzed (MRJMDAYS) is a categorical variable with days grouped into four levels (1=1-2 days, 2=3-5 days, 3=6-19 days, 4=20+ days). Because SUDAAN now needs to estimate percentages and SEs for each level of the variable instead of computing only one estimate for the variable overall, the CATLEVEL statement is introduced, and the PERCENT and SEPERCENT keywords replace the MEAN and SEMEAN keywords. The suppression rule for percentages is the same as the suppression rule for means shown in [Exhibit A.2.3](#), except PERCENT and SEPERCENT have to be divided by 100 (and thus are equivalent to MEAN and SEMEAN in the formulas). The same would apply to the means output in SAS and percentages output in R that come out as percentages and would need to be divided by 100 before applying the suppression rule. In Stata, the output will be proportions that can be directly used in the suppression rule formulas. However, if for reporting purposes, percentages need to be shown, then these proportions would need to be multiplied by 100.

### A.1.11 Testing between Overlapping Domains

[Exhibit A.2.10](#), [Exhibit A.3.10](#), [Exhibit A.4.10](#), and [Exhibit A.5.10](#) demonstrate testing between two overlapping domains, as explained in Section 7.2. This is different from the tests of between-year differences shown in [Exhibit A.2.4](#), [Exhibit A.3.4](#), [Exhibit A.4.4](#), and [Exhibit A.5.4](#). Specifically, these exhibits show how to use a stacked dataset to test whether past month

cigarette use among the full population aged 18 or older is different from cigarette use among adults aged 18 or older who are employed full time.

This code will apply both when one domain is completely contained in another and when there is only partial overlap. These exhibits use two domains, where one domain is completely contained in the other (i.e., comparing full-time employed adults with “all” adults—the employed group is completely contained by the “all” adults group). This test accounts for the correlations between the two estimates (i.e., correlation between past month cigarette use among adults aged 18 or older and past month cigarette use among adults aged 18 or older employed full time).

### **A.1.12 Testing Independence of Two Variables when One Variable Has Three or More Levels**

When comparing population subgroups defined by three or more levels of a categorical variable, log-linear chi-square tests of independence of the subgroup and the prevalence variables are conducted first to control the error level for multiple comparisons (i.e., if the goal is to compare cigarette use among several levels of employment, first test whether cigarette use is associated with employment). See Section 7.2.1 for more information. [Exhibit A.2.11](#), [Exhibit A.3.11](#), [Exhibit A.4.11](#), and [Exhibit A.5.11](#) show the code for calculating the Wald  $F$  test to determine whether cigarette use is associated with employment status. If Shah’s Wald  $F$  test (transformed from the standard Wald chi-square) indicated overall significant differences, the significance of each particular pairwise comparison of interest can be tested using the SUDAAN procedure DESCRIPT (as shown in [Exhibit A.2.10](#)), Stata ([Exhibit A.3.10](#)), SAS ([Exhibit A.4.10](#)), or R ([Exhibit A.5.10](#)). The additional pairwise testing can determine which levels of employment status show significant differences in cigarette use compared with other levels of employment as shown using SUDAAN ([Exhibit A.2.12](#)), Stata ([Exhibit A.3.12](#)), SAS ([Exhibit A.4.12](#)), or R ([Exhibit A.5.12](#)).

### **A.1.13 Testing of Linear Trends**

In addition to comparing subpopulations for one year versus another year, it can also be useful to test the linear trend for all data points across all years of interest. Linear trend testing can inform users about whether prevalence use has decreased, increased, or remained steady over the entire span of the years of interest or about changes in specific measures. [Exhibit A.2.13](#), [Exhibit A.3.13](#), [Exhibit A.4.13](#), and [Exhibit A.5.13](#) perform linear trend testing based on orthogonal polynomial coefficients. In SUDAAN, the  $t$  test method is implemented using the CONTRAST statement in the DESCRIPT procedure, as shown in [Exhibit A.2.13](#). The corresponding Stata code using test statements is shown in [Exhibit A.3.13](#), the SAS code is shown in [Exhibit A.4.13](#), and the corresponding R code is shown in [Exhibit A.5.13](#).

For each additional year of data to be included in the trend test, an additional contrast is required for that year. [Table A.3](#) is provided to assist in developing an appropriate CONTRAST statement for their desired analysis. Measures should be considered comparable across the years of interest before doing a linear trend test. When there is a trend break in the data, meaningful interpretation of analysis results may not be possible. Linear trend testing should be used when there are more than 2 years of comparable data; 4 years or more of comparable data are recommended.

## A.2 SUDAAN Exhibits

The SUDAAN exhibits in this section provide guidance on how to use various programming options to produce estimates for NSDUH using the statistical procedures documented within this report.

### Exhibit A.2.1 SUDAAN DESCRIPT Procedure (Estimate Generation: Single Year and Pooled Years of Data)

This exhibit demonstrates how to compute various types of estimates for past month alcohol (variable ALCMON) use by year and sex (variable IRSEX) for single- or combined-year (pooled) data using the SUDAAN DESCRIPT procedure. Code is included to compute the prevalence estimate (MEAN), SE of the mean (SEMEAN), weighted sample size (WSUM), unweighted sample size (NSUM), weighted total (TOTAL), and SE of the total (SETOTAL). The following optional additions to the code are displayed in the example:

- The formats defined in the OUTPUT statement increase 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 for the user.
- Design effects for each mean estimate can be output by specifying the DEFFMEAN option in the OUTPUT statement.

Whether the SETOTAL is taken directly from SUDAAN depends on whether the specified domain (i.e., sex in this example) is fixed (i.e., domains forced to match their respective U.S. Census Bureau or ACS population estimates through the weight calibration process). For additional information on SEs, see Section A.1.5. For more information on how to create a pooled weight to use when producing annual averages of combined years of data, see Chapter 3. The input dataset would need to include at least 2 years of NSDUH data with a YEAR variable defined for each year on the dataset.

```
PROC SORT DATA=DATANAME; /*SAS code to sort output dataset with 2
years of NSDUH data by Nesting Variables*/
BY VESTR VEREP;
RUN;

PROC DESCRIPT DATA=DATANAME DDF=750 DESIGN=WR FILETYPE=SAS DEFT4;
/*Alternatively, the DDF may change if using combined data based
on the analysis being performed; see Table 6.3 in Chapter 6*/
NEST VESTR VEREP;
WEIGHT ANALWT2; /*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.*/
VAR ALCMON; /*Past month alcohol analysis variable*/
SUBGROUP YEAR IRSEX;
    /*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*/
    /*Sex variable, where male=1 & female=2*/
LEVELS 2 2; /* For each variable listed on the subgroup
statement, list the number of categorical levels */
```

### Exhibit A.2.1 SUDAAN DESCRIPT Procedure (Estimate Generation: Single Year and Pooled Years of Data) (continued)

```
TABLES YEAR*IRSEX; /*Sex by year*/
PRINT WSUM NSUM MEAN SEMEAN TOTAL SETOTAL / REPLACE STYLE=NCHS;
OUTPUT WSUM MEAN SEMEAN TOTAL SETOTAL NSUM 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";
TITLE "ESTIMATES OF PAST MONTH ALCOHOL BY YEAR AND SEX";
RUN;
```

The following CLASS statement could be used in place of SUBGROUP and LEVELS statements in the above example. 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. Using either of these methods differs from subsetting the population to a specific domain, which is shown in [Exhibit A.2.7](#).

```
CLASS YEAR IRSEX;
```

### Exhibit A.2.2 SAS Code Based on SUDAAN Output (Calculation of Standard Error of Totals for Fixed Domains)

This exhibit computes the SE of the total for fixed domains using the data produced by the example in [Exhibit A.2.1](#). Because sex is a fixed domain, the SE of the totals would not be taken directly from the example in [Exhibit A.2.1](#) but rather would be computed using the alternative SE estimation method as shown below.

```
DATA ESTIMATE;
SET OUT.SUDFILE; /* input the output file from Exhibit A.2.1
SUDAAN procedure */

/*****
Define SETOTAL for sex because it is listed as a fixed domain in
Table 5.1.
In the SUDAAN procedure in Exhibit A.2.1, IRSEX is in the
subgroup Statement with 2 levels indicated. Therefore,
values for 0=total male & females, 1=males, and 2=females
are automatically produced in SUDAAN. Because all three of
these groups (i.e., total, males, and females) are fixed
domains for ages 12 or older, the alternative calculation
of the SE of totals performed in SAS is applied to three
levels.
*****/

IF IRSEX IN (0,1,2) THEN SETOTAL=WSUM*SEMEAN;

RUN;
```

### Exhibit A.2.3 SAS Code Based on SUDAAN Output (Implementation of Suppression Rule)

This exhibit applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1 (commented out in the exhibit) using the data produced in [Exhibit A.2.1](#). Starting with the 2024 NSDUH, the suppression rule was simplified, and the code for the new rules described in [Table 9.1](#) is shown in this exhibit.

```
DATA ESTIMATE;
SET OUT.SUDFILE; /*input the output file from Exhibit A.2.1
SUDAAN procedure*/

/*****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)));

/*SUPPRESSION RULE FOR PREVALENCE ESTIMATES*/
IF (SEMEAN=0) OR (RSELNP GT 0.175) OR (NSUM < 50) 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;
```

### Exhibit A.2.4 SUDAAN DESCRIPT Procedure (Tests of Differences)

This exhibit performs significance testing between comparable years of NSDUH data using the data produced by the example in [Exhibit A.2.1](#). The input dataset includes 2 years of NSDUH data sorted by nesting variables. 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 between the 2 years, as specified in the DIFFVAR statement.

```
PROC DESCRIPT DATA=DATANAME DDF=750 DESIGN=WR FILETYPE=SAS;
NEST VESTR VEREP;
WEIGHT ANALWT2; /*Standard single-year, person-level analysis
weight*/
VAR ALCMON; /*Past month alcohol analysis variable*/
SUBGROUP YEAR IRSEX;
LEVELS 2 2;
TABLES IRSEX;
DIFFVAR YEAR=(1 2) / NAME="YEAR1 vs YEAR2";
PRINT WSUM NSUM MEAN SEMEAN TOTAL SETOTAL T_MEAN P_MEAN /
REPLACE STYLE=NCHS;
OUTPUT WSUM MEAN SEMEAN TOTAL SETOTAL NSUM T_MEAN P_MEAN /
REPLACE
```

**Exhibit A.2.4 SUDAAN DESCRIPT Procedure (Tests of Differences) (continued)**

```

NSUMFMT=F8.0 WSUMFMT=F12.0 MEANFMT=F15.10 SEMEANFMT=F15.10
TOTALFMT=F12.0 SETOTALFMT=F12.0 FILENAME="OUT.SUDTESTS";
TITLE "TESTS OF DIFFERENCES BETWEEN YEAR1 AND YEAR2 ESTIMATES OF
PAST MONTH ALCOHOL BY SEX";
RUN;

```

Note: For testing of multiple pairs of years, more years could be included in the dataset (and LEVELS statement), and several DIFFVAR statements could be used in place of the single DIFFVAR statement in the above example. For example, for a dataset with 4 years:

```

LEVELS 4 2;
DIFFVAR YEAR=(1 4) /NAME="YEAR1 vs YEAR 4";
...
DIFFVAR YEAR=(3 4) /NAME="YEAR3 vs YEAR4";
TITLE "TESTS OF DIFFERENCES BETWEEN MULTIPLE YEARS OF PAST MONTH
ALCOHOL USE BY SEX";

```

When one or more contrasts are specified in SUDAAN, as shown using the DIFFVAR statement above, the output variable MEAN becomes the contrast mean where the number assigned to the output variable, CONTRAST, represents the tests in order of appearance in the SUDAAN code, and SEMEAN becomes the SE of the contrast mean.

SUDAAN does not test differences in the corresponding totals explicitly. However, it will output the contrast total (TOTAL) and the SE of the contrast total (SETOTAL).

**Exhibit A.2.5 SAS Code Based on SUDAAN Output (Calculation of the *P* Value for the Test of Differences between Totals for Nonfixed Domains)**

With the statistics and the correct degrees of freedom (750 in this example), the *p* value (PVALT) for the test of differences between totals for nonfixed domains can be calculated as shown below. The SAS function PROBT returns the probability from a *t*-distribution.

```

DATA ESTIMATE;
SET OUT.SUDTESTS; /*input the output file from Exhibit A.2.4
SUDAAN procedure*/
IF SETOTAL GT 0.0 THEN DO;
PVALT=2*(1-PROBT(ABS(TOTAL/SETOTAL),750));
END;
RUN;

```

**Exhibit A.2.6 SUDAAN DESCRIPT Procedure and SAS Code Based on SUDAAN Output (Covariance Matrix, Identification of Covariance Components, Calculation of Variances, and Calculation of the *P* Value for the Test of Differences between Totals for Fixed Domains)**

To calculate the *p* value for the test of differences between totals for fixed domains, three SAS datasets, one containing the covariances (COV) and two containing the variances (EST1 and EST2), are then merged with the output dataset (SUDTESTS) from the DESCRIPT procedure that generated the tests of differences in [Exhibit A.2.4](#).

**Exhibit A.2.6 SUDAAN DESCRIPT Procedure and SAS Code Based on SUDAAN Output (Covariance Matrix, Identification of Covariance Components, Calculation of Variances, and Calculation of the *P* Value for the Test of Differences between Totals for Fixed Domains) (continued)**

The covariances of the estimated means can be obtained from the output of the DESCRIPT procedure shown below. The covariance matrix in SUDAAN consists of a row and column for each sex (total, male, female) and year (both years; i.e., YEAR1 and YEAR2) combination with each cell corresponding to a particular variance component (i.e., a 9 × 9 matrix). Because the rows and columns of the matrix are identical, the cells in the top half (above the diagonal) and the bottom half (below the diagonal) are identical. [Table A.2](#) shows a shell for what the SUDAAN covariance matrix would look like for this example.

**Table A.2 SUDAAN Matrix Shell**

IRSEX	YEAR	ROWNUM	IRSEX=0			IRSEX=1			IRSEX=2		
			YEAR=0	YEAR=1	YEAR=2	YEAR=0	YEAR=1	YEAR=2	YEAR=0	YEAR=1	YEAR=2
			B01	B02	B03	B04	B05	B06	B07	B08	B09
IRSEX=0	YEAR=0	1									
	YEAR=1	2									
	YEAR=2	3									
IRSEX=1	YEAR=0	4									
	YEAR=1	5									
	YEAR=2	6									
IRSEX=2	YEAR=0	7									
	YEAR=1	8									
	YEAR=2	9									

In the SUDAAN output, each cell of the variance-covariance matrix is identified by a separate variable of the form B0*x*, where *x* is a particular cell number. (Cells are numbered left to right.) The variable *ROWNUM* is an additional output variable that simply identifies the matrix row. The covariance data needed for a particular significance test can be pulled out of the matrix using SAS code. For this example, the covariance for IRSEX=0 between YEAR=1 and YEAR=2 would be B03 from ROWNUM2 or B02 from ROWNUM3. These two values would be the same in this case. The needed covariances are kept in the SAS code shown that creates dataset COV.

```
PROC DESCRIPT DATA=DATANAME DDF=750 DESIGN=WR FILETYPE=SAS DEFT4;
  NEST VESTR VEREP;
  WEIGHT ANALWT2; /*Standard single-year, person-level analysis
  weight*/
  VAR ALCMON; /*Past month alcohol analysis variable*/
  SUBGROUP YEAR IRSEX;
  LEVELS 2 2;
  TABLES IRSEX*YEAR;
  PRINT COVMEAN / STYLE = NCHS;
  OUTPUT / MEANCOV = DEFAULT REPLACE FILENAME= "OUT.SUDCOV";
  TITLE "Variance Covariance Matrices";
  RUN;
```

### Exhibit A.2.6 SUDAAN DESCRIPT Procedure and SAS Code Based on SUDAAN Output (Covariance Matrix, Identification of Covariance Components, Calculation of Variances, and Calculation of the $P$ Value for the Test of Differences between Totals for Fixed Domains) (continued)

```
DATA COV(KEEP=IRSEX COV1);
SET OUT.SUDCOV01;
IF ROWNUM=2 THEN DO; IRSEX=0; COV1=B03; END;
ELSE IF ROWNUM=8 THEN DO; IRSEX=2; COV1=B09; END;
ELSE IF ROWNUM=5 THEN DO; IRSEX=1; COV1=B06; END;

IF ROWNUM IN (2,5,8) THEN OUTPUT;

RUN;

PROC SORT DATA=COV;
BY IRSEX;
RUN;
```

The variances of the means are estimated in the following separate data steps and output into datasets EST1 and EST2. The variance is the square of the SE of the mean. The SEs of the means were output in the original procedure that generated the estimates in [Exhibit A.2.1](#).

```
DATA EST1(KEEP=WSUM1 VAR1 YEAR IRSEX);
SET OUT.SUDFILE;
WHERE YEAR=1;
WSUM1=WSUM;
VAR1=SEMEAN**2; /*THE variance is the SEMEAN squared*/
RUN;

DATA EST2(KEEP=WSUM2 VAR2 YEAR IRSEX);
SET OUT.SUDFILE;
WHERE YEAR=2;
WSUM2=WSUM;
VAR2 = SEMEAN**2;
RUN;
```

With the proper statistics contained in one dataset (data P\_VALUE), the corresponding  $p$  value for the tests of differences between fixed domain totals can be produced using the SAS PROBT function and calculated  $t$  test statistic as shown below.

```
DATA P_VALUE;
MERGE EST1 EST2 OUT.SUDTESTS COV;
BY IRSEX;

PVALT=2*(1-PROBT(ABS(TOTAL/SQRT(WSUM1**2*VAR1+WSUM2**2*VAR2-
2*WSUM1*WSUM2*COV1)),750));
RUN;
```

**Exhibit A.2.7 SAS Code (Recoding a Variable) and SUDAAN DESCRIPT Procedure (Estimate Generation with (1) Missing Values and (2) Using Subpopulation)**

This exhibit shows two methods for handling unused values in variable recodes in estimate generation using a single year of NSDUH data. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```

PROC SORT DATA=DATANAME_SINGLE;
BY VESTR VEREP;
RUN;

/* Method 1, recoding unused values to SAS missing*/

DATA RECODES;
SET DATANAME_SINGLE;
IF IRMJAGE=991 THEN IRMJAGE_R=.;
ELSE IRMJAGE_R=IRMJAGE;
RUN;

PROC DESCRIPT DATA=RECODES DDF=750 DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR VEREP;
WEIGHT ANALWT2; /*Standard single-year, person-level analysis
weight*/
VAR IRMJAGE_R; /*Marijuana Age of First Use recoded analysis
variable*/
SUBGROUP IRSEX; /*Sex variable, where male=1 & female=2*/
LEVELS 2;
TABLES IRSEX; /*Sex*/

PRINT MEAN SEMEAN / REPLACE STYLE=NCHS;
TITLE "ESTIMATES OF AGE OF FIRST USE OF MARIJUANA BY SEX";
RUN;

/* Method 2, using subpopulation statement in SUDAAN to omit the
unused values*/

PROC DESCRIPT DATA=DATANAME_SINGLE DDF=750 DESIGN=WR FILETYPE=SAS
DEFT4;
NEST VESTR VEREP;
WEIGHT ANALWT2; /*Standard single-year, person-level analysis
weight*/
SUBPOPN MRJFLAG=1; /*Subsetting to omit those respondents who had
never used marijuana, i.e., omitting respondents where
IRMJAGE=991*/
VAR IRMJAGE; /*Marijuana Age of First Use analysis variable*/
SUBGROUP IRSEX; /*Sex variable, where male=1 & female=2*/
LEVELS 2;
TABLES IRSEX; /*Sex*/
PRINT MEAN SEMEAN / REPLACE STYLE=NCHS;
TITLE "ESTIMATES OF AGE OF FIRST USE OF MARIJUANA BY SEX";
RUN;

```

**Exhibit A.2.8 SAS Code Based on SUDAAN Output (Calculating a 95 Percent Confidence Interval)**

This SAS code computes the CIs using the output data from [Exhibit A.2.1](#).

```
DATA CI;
SET OUT.SUDFILE; /*output data from Exhibit A.2.1*/
T_QNTILE=TINV(0.975,750); /*define t-statistic*/
NUMBER=SEMEAN/(MEAN*(1-MEAN));
L=LOG(MEAN/(1-MEAN));

A=L-T_QNTILE*NUMBER;
B=L+T_QNTILE*NUMBER;

/*FLOWER AND PUPPER ARE THE 95% CIS ASSOCIATED WITH MEAN FROM
SUDAAN*/
FLOWER=1/(1+EXP(-A));
PUPPER=1/(1+EXP(-B));

/*TLOWER AND TUPPER ARE THE 95% CIS ASSOCIATED WITH TOTAL FROM
SUDAAN*/
TLOWER=WSUM*FLOWER;
TUPPER=WSUM*PUPPER;
RUN;
```

**Exhibit A.2.9 SUDAAN DESCRIPT Procedure (Estimate Generation for Categorical Variable, i.e., Number of Days Used Substance in the Past Month among Past Month Users)**

This exhibit shows how to compute estimates corresponding to levels of a categorical variable (MRJMDAYS). DATANAME\_SINGLE is a dataset that has been subset to a single year of data. Similar to [Exhibit A.2.1](#), the weighted sample size (WSUM), unweighted sample size (NSUM), weighted total (TOTAL), and SE (SETOTAL) 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.

```
PROC DESCRIPT DATA=DATANAME_SINGLE DDF=750 DESIGN=WR FILETYPE=SAS
DEFT4;
NEST VESTR VEREP;
WEIGHT ANALWT2; /*Standard single-year, person-level analysis
weight*/
VAR MRJMDAYS MRJMDAYS MRJMDAYS MRJMDAYS; /*Marijuana Use
frequency in the past month variable: 1=1-2 days, 2=3-5 days,
3=6-19 days, 4=20+ days, 5=did not use in the past month*/
CATLEVEL 1 2 3 4; /*levels of MRJMDAYS to be shown in table*/
```

### Exhibit A.2.9 SUDAAN DESCRIPT Procedure (Estimate Generation for Categorical Variable, i.e., Number of Days Used Substance in the Past Month among Past Month Users) (continued)

```
SUBGROUP MRJMON; /*Past month marijuana use variable, where used
in past month=1 & did not use in past month=0*/
LEVELS 1;
TABLES MRJMON; /*Tables will show percentages among marijuana
users*/
PRINT WSUM NSUM PERCENT SEPERCENT TOTAL SETOTAL / REPLACE
STYLE=NCHS;
OUTPUT WSUM NSUM PERCENT SEPERCENT TOTAL SETOTAL / REPLACE
FILENAME= "OUT.SUDFILE_FREQ";
TITLE "FREQUENCY OF MARIJUANA USE BY PAST MONTH MARIJUANA USERS";
RUN;
```

### Exhibit A.2.10 SAS Code (Stacking a Dataset) and SUDAAN DESCRIPT Procedure (Test of Difference when Two Groups Overlap Using Stacked Data)

This exhibit shows how to test among overlapping domains (i.e., scenarios where some cases are in both domains are being compared). A stacked dataset is created first that includes two records for each respondent in the overlap. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```
DATA STACKED;
SET DATANAME_SINGLE(IN=A) DATANAME_SINGLE(IN=B); /*reading in
data twice*/
/*create an indicator variable for the stacked data, this will be
used in the diffvar statement in PROC DESCRIPT
When indic=1, employ=1 represents the full population
When indic=2, employ=1 represents those employed full time*/
IF A THEN DO;
INDIC=1;
IF IRWRKSTAT18 IN (1,2,3,4) THEN EMPLOY=1;
/* IRWRKSTAT18 is a four-level employment variable for adults,
where level 1 is those employed full time, 2 is those employed
part time, 3 are those unemployed, and 4 are all other adults.
Respondents aged 12 to 17 are coded as level 99*/
ELSE EMPLOY=0;
END;
ELSE IF B THEN DO;
INDIC=2;
IF IRWRKSTAT18=1 THEN EMPLOY=1;
ELSE EMPLOY=0;
END;
RUN;

PROC SORT DATA=STACKED;
BY VESTR VEREP;
RUN;
```

The stacked data are then used to compute the test of differences among the overlapping domains.

**Exhibit A.2.10 SAS Code (Stacking a Dataset) and SUDAAN DESCRIPT Procedure (Test of Difference when Two Groups Overlap Using Stacked Data) (continued)**

```

PROC DESCRIPT DATA=STACKED DDF=750 DESIGN=WR FILETYPE=SAS;
NEST VESTR VEREP;
WEIGHT ANALWT2; /*Standard single-year, person-level analysis
weight*/
VAR CIGMON; /*Past month cigarette use analysis variable*/
SUBGROUP INDIC;
LEVELS 2;
DIFFVAR INDIC=(1 2); /*Since subsetting in the next line to
employ=1, this is testing all persons 18+ vs. employed persons
18+*/
SUBPOPN CATAG18=1 AND EMPLOY=1;
PRINT WSUM NSUM MEAN SEMEAN TOTAL SETOTAL T_MEAN P_MEAN /
  REPLACE STYLE=NCHS;
OUTPUT WSUM MEAN SEMEAN TOTAL SETOTAL NSUM 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_STACKED";
TITLE "TESTS OF DIFFERENCES BETWEEN ALL PERSONS 18 OR OLDER AND
EMPLOYED PERSONS 18 OR OLDER";
RUN;

```

**Exhibit A.2.11 SUDAAN CROSSTAB Procedure (Test for Independence Based on a Log-Linear Model)**

A two-step process is needed when comparing population subgroups defined by three or more levels of a categorical variable. The first step shown in this exhibit is to create a log-linear chi-square test of independence of the subgroup and the prevalence variables to control the error level for multiple comparisons. This is done by calculating Shah's Wald  $F$  test, which indicates whether a statically significant association exists between past month cigarette use and employment status. Performing statistical tests in SUDAAN requires the specification of degrees of freedom using the DDF option. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```

PROC CROSSTAB DATA=DATANAME_SINGLE DDF=750 DESIGN=WR FILETYPE=SAS
DEFT4;
NEST VESTR VEREP;
WEIGHT ANALWT2; /*Standard single-year, person-level analysis
weight*/
CLASS CIGMON; /*Past month cigarette use analysis variable*/
SUBGROUP IRWRKSTAT18; /*four level employment status variable*/
LEVELS 4;
TABLES IRWRKSTAT18*CIGMON;
TEST LLCHISQ / WALDF; /*log linear hypothesis test, wald F test
statistic, if test statistic is significant, then reject null
hypothesis of no interaction*/
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;

```

**Exhibit A.2.12 SUDAAN DESCRIPT Procedure (Pairwise Testing)**

Once an association is determined by a significant Wald  $F$  test computed in [Exhibit A.2.11](#), then pairwise testing as shown below can be done to determine individual significant tests across levels of an independent variable (e.g., levels of employment status). Values of T\_MEAN and P\_MEAN represent the  $t$  test statistic and  $p$  value, respectively, for each  $k(k-1)/2$  contrast, where  $k$  is the number of levels of the PAIRWISE statement variable. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```
PROC DESCRIPT DATA=DATANAME_SINGLE DDF=750 DESIGN=WR
FILETYPE=SAS;
NEST VESTR VEREP;
WEIGHT ANALWT2; /*Standard single-year, person-level analysis
weight*/
VAR CIGMON; /*Past month cigarette use analysis variable*/
SUBGROUP IRWRKSTAT18; /*four level employment status variable*/
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;
OUTPUT WSUM MEAN SEMEAN TOTAL SETOTAL NSUM 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_PAIRWISE";
TITLE "TESTS OF DIFFERENCES IN PAST MONTH CIGARETTE USE AMONG ALL
LEVELS OF EMPLOYMENT STATUS";
RUN;
```

**Exhibit A.2.13 SUDAAN DESCRIPT Procedure (Test of Linear Trends with DESCRIPT)**

This exhibit shows how to do a linear trend test to inform users about whether prevalence use has decreased, increased, or remained steady over the entire span of the years of interest. This exhibit shows tests for a linear trend in past month alcohol use (variable ALCMON) by sex across 4 years (variable YEAR) to evaluate change over time. Various methods can be used to test a linear trend. However, this example uses the SUDAAN procedure DESCRIPT with a CONTRAST statement, a nonparametric method consistent with what is used in the production of NSDUH data products. Other types of examples using modeling can be found in the *2024 National Survey on Drug Use and Health: Public Use File Data Users' Guide* (CBHSQ, 2025e). To use the contrast totals or SEs of totals, the analysis weight must be modified as discussed in Chapter 3 to produce accurate estimates. This modification is not necessary if only the  $t$  test statistic and the  $p$  value are of interest. DATANAME is a dataset that has 4 years of data with a year variable defined for each year on the dataset.

```
PROC DESCRIPT DATA=DATANAME DDF=750 DESIGN=WR FILETYPE=SAS;
NEST VESTR VEREP;
WEIGHT ANALWT2;
VAR ALCMON;
```

**Exhibit A.2.13 SUDAAN DESCRIPT Procedure (Test of Linear Trends with DESCRIPT) (continued)**

```

SUBGROUP YEAR IRSEX;
LEVELS 4 2;
TABLES IRSEX;
CONTRAST YEAR = (-3 -1 1 3) / NAME="4 YEAR LINEAR TREND TEST";
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="SUDTESTS_LINEAR";
TITLE "TEST OF LINEAR TREND IN PAST MONTH ALCOHOL USE BY SEX";
RUN;

```

The number of coefficients specified in the CONTRAST vector must match the number of years of data a user wants to include in the trend test. [Table A.3](#) is provided to help develop an appropriate CONTRAST statement for the desired analysis. Measures should be considered comparable across the years of interest before doing a linear trend test. When there is a trend break in the data, meaningful interpretation of analysis results may not be possible.

**Table A.3 Contrast Statements for Nonparametric Linear Trend Testing**

Number of Years	CONTRAST Statement
12	(-11 -9 -7 -5 -3 -1 1 3 5 7 9 11)
11	(-5 -4 -3 -2 -1 0 1 2 3 4 5)
10	(-9 -7 -5 -3 -1 1 3 5 7 9)
9	(-4 -3 -2 -1 0 1 2 3 4)
8	(-7 -5 -3 -1 1 3 5 7)
7	(-3 -2 -1 0 1 2 3)
6	(-5 -3 -1 1 3 5)
5	(-2 -1 0 1 2)
4	(-3 -1 1 3)

## A.3 Stata Exhibits

The Stata exhibits in this section provide guidance on how to use various programming options to produce estimates for NSDUH using the statistical procedures documented within this report. Note that Stata code is case sensitive. Stata does not require sorting by the nesting variables before proceeding with the analysis.

### Exhibit A.3.1 Stata COMMANDS `svy: mean` and `svy: total` (Estimate Generation: Single Year and Pooled Years of Data)

This exhibit 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, SE of the mean, weighted sample size, unweighted sample size, design effect, weighted total, and SE of the totals. 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 (i.e., domains forced to match their respective U.S. Census Bureau and ACS population estimates through the weight calibration process). For additional information on SEs, see Section A.1.5. For more information on how to create a pooled weight to use when producing annual averages of combined years of data, see Chapter 3. `DATANAME_SINGLE` is a dataset that has been subset to a single year of data. If running pooled data estimates, then the code would need to be adjusted, and the input dataset would need to include at least 2 years of NSDUH data.

```
use using ".\\dataname_single.dta", clear

/*Ensure all variables are lower case*/
rename *, lower

/*ID Nesting variables (VESTR and VEREP) and weight variable
(ANALWT2 - 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. The DOF
may also change if using combined data depending on the analysis
being performed; see Table 6.3 in Chapter 6*/
svyset verep [pw=analwt2], strata(vestr) dof(750)

gen double total_out=.
gen setotal=.
gen mean_out=.
gen semean=.
gen nsum=.
gen double wsum=.
gen deffmean=.

/*Estimated means of past month alcohol use by sex*/

/*Sex variable, where male=1 & female=2*/
svy: mean alcmon, over(irsex)
matrix M=e(b) /*Store mean estimates in matrix M*/
matrix S=e(V) /*Store variances in matrix S*/
matrix N=e(_N) /*Store sample size in matrix N*/
```

**Exhibit A.3.1 Stata COMMANDS svy: mean and svy: total (Estimate Generation: Single Year and Pooled Years of Data) (continued)**

```

matrix W=e(_N_subp) /*Store weighted sample size in matrix W*/
estat effects, deff srssubpop/*Obtain design effect*/
matrix D=e(deff) /*Store design effect in matrix D*/

/*Extract values stored in the M, S, N, W, and D matrices defined
above to the mean_out, semean, nsum, wsum, and deffmean
variables. The loop ensures that the appropriate values are
extracted for each value of sex.*/

    forvalues j=1/2 { /* number of sex categories*/
        replace mean_out=(M[1,`j']) if irsex==`j'
    edi    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'
    }

/*Estimated Totals*/
svy: total alcmon, over(irsex)

matrix M=e(b) /*Store total estimates in matrix M*/
matrix S=e(V) /*Store variances in matrix S*/

/*Extract values stored in the M and S matrices defined above to
the total_out and setotal variables. The loop ensures that the
appropriate values are extracted for value of year and sex.*/

    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 /*keep one record per subpopulation
of interest*/

/*Format wsum, mean_out, semean, total_out, setotal, nsum, and
deffmean variables to control appearance in output.*/

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

/*Estimates of past month alcohol by sex*/
list irsex wsum nsum mean_out semean total_out setotal
    
```

**Exhibit A.3.1 Stata COMMANDS svy: mean and svy: total (Estimate Generation: Single Year and Pooled Years of Data) (continued)**

*/\*The output from this exhibit will be utilized in [Exhibit A.3.6](#). Users can either rerun the code presented in this exhibit or save the output from this exhibit to a dataset using the following command.\*/*

```
save ".\\EXmean.dta" , replace
```

To compute estimates corresponding to levels of a categorical variable, the following code provides general details that could be applied to the code above:

```
use using ".\\dataname.dta", clear
rename *, lower
svyset verep [pw=analwt2], strata(vestr) dof(750)
svy: proportion mrjmdays, subpop(mrjmon)
```

**Exhibit A.3.2 Stata Code (Calculation of Standard Error of Totals for Fixed Domains)**

This exhibit computes the SE of the total for fixed domains using the data produced by the example in [Exhibit A.3.1](#). Because sex is a fixed domain, the SE of the totals would not be taken directly from the example in [Exhibit A.3.1](#) but rather would be computed using the alternative SE estimation method as shown below. Unlike SUDAAN, Stata does not automatically produce overall estimates. Therefore, the alternative calculation is applied to only the two levels of the IRSEX variable.

```
generate setotal2=wsum*semean
replace setotal = setotal2 if inlist(irsex,1,2)
/*Note, Stata does not automatically produce overall estimates,
i.e., irsex=0*/
```

**Exhibit A.3.3 Stata Code (Implementation of Suppression Rule)**

This exhibit applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1 (commented out in the exhibit) using the data produced by the example in [Exhibit A.3.1](#). Starting with the 2024 NSDUH, the suppression rule was simplified and the code for the new rules described in [Table 9.1](#) is in this exhibit.

```
/******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
```

**Exhibit A.3.3 Stata Code (Implementation of Suppression Rule) (continued)**

```

/*SUPPRESSION RULE FOR PREVALENCE ESTIMATES*/
generate suprul1=1 if semean == 0 & !missing(semean)
generate suprul2=1 if rselnp > 0.175 & !missing(rselnp)
generate suprul3=1 if nsum < 50 & !missing(nsum)

generate suppress=0
replace suppress=1 if suprul1==1 | /// suprul2==1 | suprul3==1

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

```

**Exhibit A.3.4 Stata COMMANDS svy: mean and svy: total (Tests of Differences)**

This exhibit performs significance testing between comparable years of NSDUH data. The input dataset would need to include at least 2 years of NSDUH data with a YEAR variable defined for each year on the dataset. Creation of this variable has been omitted from the following example, which estimates differences in mean past month alcohol use (variable ALCMON) between 2 survey years. Stata does not require the data to be sorted by nesting variables.

```

use using ".\\dataname.dta", clear

/*Ensure all variables are lower case*/
rename *, lower

/*ID Nesting variables (VESTR and VEREP) and weight variable
(ANALWT2 - standard single-year, person-level analysis weight)*/
svyset verep [pweight=analwt2], strata(vestr) dof(750)
{
svy: mean alcmon, over(year)

matrix Me = e(b)

local max=2 *number of years
/*Define the matrix output with J(x,7,.) where x is the number of
comparisons*/
matrix output = J(1,7,.)

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
    }
}

```

### Exhibit A.3.4 Stata COMMANDS svy: mean and svy: total (Tests of Differences) (continued)

```

        test c.alcmon@`j'.year = c.alcmon@`stop'.year,
        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 alcmon, over(year)

matrix M = e(b)
local max=2

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 c.alcmon@`j'.year = c.alcmon@`stop'.year,
        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 setotal
mean_pval

matrix list output

```

For testing of more than 2 years of data, 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 pairwise difference of years requires the following code changes to be made:

```

local max=3 /*3 years*/

matrix output = J(3,7,.) /*3 comparisons (1 vs 2, 1 vs 3, and 2
vs 3)*/

```

### Exhibit A.3.5 Stata Code (Calculation of the $P$ Value for the Test of Differences between Totals for Nonfixed Domains)

With just a few modifications to the code from [Exhibit A.3.4](#), the  $p$  value (PVALT) can be calculated for the test of differences between totals for nonfixed domains.

```
/*Change the dimensions of the output matrix*/
/*Add row to the output matrix*/
matrix output = J(1,8,.)
/*Add the line below to the second set of For Loops*/
matrix output [`contrast',8]=r(p)
/*Add (PVALT) to the list of column names*/
matrix colnames output = level1 level2 mean semean total setotal
mean_pval pvalt
```

### Exhibit A.3.6 Stata COMMAND svy: mean (Covariance Matrix, Identification of Covariance Components, Calculation of Variances, Calculation of the $P$ Value for the Test of Differences between Totals for Fixed Domains)

To calculate the  $p$  value for the test of differences between totals for fixed domains, the covariance matrix and weighted counts matrix from the SVY procedure need to be saved. The weighted counts and variances of the 2 compared years, along with the covariance between the 2 years, need to be added to the output matrix from [Exhibit A.3.4](#). The input dataset would need to include at least 2 years of NSDUH data with a YEAR variable defined for each year on the dataset.

[Table A.4](#) presents the Stata matrix shell.

**Table A.4 Stata Matrix Shell**

Subpopulation	c.alcmon@1.year	c.alcmon@2.year
c.alcmon@1.year		
c.alcmon@2.year		

```
use using ".\\dataname.dta", clear
/*Ensure all variables are lower case*/
rename *, lower

/*ID Nesting variables (VESTR and VEREP) and weight variable
(ANALWT2 - standard single-year, person-level analysis weight)*/

svyset verep [pweight=analwt2], strata(vestr) dof(750)
svy: mean alcmon, over(year)

*Save and display the Covariance Matrix
matrix V = e(V)
```

**Exhibit A.3.6 Stata COMMAND svy: mean (Covariance Matrix, Identification of Covariance Components, Calculation of Variances, Calculation of the P Value for the Test of Differences between Totals for Fixed Domains) (continued)**

```

matrix list V

local max=2      /*number of years*/

matrix output = J(1,9,.) /*number of contrasts*/
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 c.alcmon@`j'.year = c.alcmon@`stop'.year, nosvyadjust
        matvlc(mtest`contrast')
            matrix output[`contrast',7]=V[`j',`j']
            matrix output[`contrast',8]=V[`stop',`stop']
            matrix output[`contrast',9]=V[`j',`stop']
        }
        local counter2 = `counter2' - 1
    }

svy: total alcmon, over(year)

matrix W=e(_N_subp)
matrix M = e(b)
local max=3

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 c.alcmon@`j'.year = c.alcmon@`stop'.year, nosvyadjust
        matvlc(test`contrast')
            matrix output[`contrast', 1]=`j'
            matrix output[`contrast', 2]=`stop'
            matrix output[`contrast',5]=W[1,`j']
            matrix output[`contrast',6]=W[1,`stop']
            matrix output[`contrast',3]=M[1,`j']-M[1,`stop']
            matrix output[`contrast',4]=
tprob(750,abs(output[`contrast',3]/sqrt(output[`contrast',5]^2 *
output[`contrast',7] + output[`contrast',6]^2 *
output[`contrast',8] - 2 * output[`contrast',5] *
output[`contrast',6] * output[`contrast',9])))
        }
    }

```

**Exhibit A.3.6 Stata COMMAND svy: mean (Covariance Matrix, Identification of Covariance Components, Calculation of Variances, Calculation of the P Value for the Test of Differences between Totals for Fixed Domains) (continued)**

```

local counter2 = `counter2' - 1
}
matrix colnames output = level1 level2 total pvalt wsum1 wsum2
var1 var2 cov1
matrix list output

```

**Exhibit A.3.7 Stata Code (Recoding a Variable, Estimate Generation with (1) Missing Values and (2) Using Subpopulation)**

This exhibit shows two methods for handling unused values in variable recodes in estimate generation using a single year of NSDUH data. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```

/*Read in data*/
use using ".\\dataname_single.dta", clear
/*Ensure all variables are lower case*/
rename *, lower

/*Method 1, recoding unused values to missing*/
generate irmjage_r = irmjage
replace irmjage_r = . if irmjage == 991
svyset verep [pweight=analwt2], strata(vestr) dof(750)
svy: mean irmjage_r, over(irsex)
/*marijuana age of first use analysis variable, sex variable*/

/*Method 2, using subpopulation to omit the unused values*/
svyset verep [pweight=analwt2], strata(vestr) dof(750)
generate subpop=1 if irmjage != 991

svy, subpop(subpop): mean irmjage, over(irsex)

```

**Exhibit A.3.8 Stata Code (Calculating a 95 Percent Confidence Interval for a Mean)**

This exhibit computes a 95 percent CI by using the output data from [Exhibit A.3.1](#).

```

/*Run code from Exhibit A.3.1 or save output dataset from
Exhibit A.3.1 and use that as input to this code.*/
generate t_qntile = invt(750,0.975)
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))

/*plower and pupper are the 95% CIs associated with mean_out from
Stata*/

generate tlower = wsum*plower

```

**Exhibit A.3.8 Stata Code (Calculating a 95 Percent Confidence Interval for a Mean) (continued)**

```

generate tupper = wsum*pupper

/*tlower and tupper are the 95% CIs associated with total_out
from Stata*/
duplicates drop year irsex, force /*keep one record per
subpopulation of interest*/

keep year irsex nsum wsum mean_out semean total_out setotal
///t_qntile number 1 a b plower pupper tlower tupper

```

**Exhibit A.3.9 Stata Code (Estimate Generation for Categorical Variable; i.e., Number of Days Used Substance in the Past Month among Past Month Users)**

This exhibit shows how to compute estimates corresponding to levels of a categorical variable. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```

use using ".\dataname_single.dta", clear
/*Ensure all variables are lower case*/
rename *, lower

svyset verep [pw=analwt2], strata(vestr) dof(750)
svy: proportion mrjmdays, subpop(mrjmon)
/*This code will produce output showing proportions for marijuana
use frequency in the past month, to get percentages, these proportions
would need to be multiplied by 100*/

```

**Exhibit A.3.10 Stata Code (Test of Difference when Two Groups Overlap Using Stacked Data)**

This exhibit shows how to test among overlapping domains (i.e., scenarios where some cases are in both domains are being compared). A stacked dataset is created first that includes two records for each respondent in the overlap. The stacked data are then used to compute the test of differences among the overlapping domains. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```

/*create an indicator variable for the stacked data, this will be
used in the mean statement within the svy command
When indic=1, employ=1 represents the full population
When indic=2, employ=1 represents those employed full time*/

/*Creating the first dataset*/
/*Read in data */
use using ".\dataname_single.dta", clear
gen indic = 1

/*Stack on the second dataset*/
append using ".\dataname_single.dta"
replace indic = 2 if indic == .

```

**Exhibit A.3.10 Stata Code (Test of Difference when Two Groups Overlap Using Stacked Data) (continued)**

```

/*Ensure all variables are lower case*/
rename *, lower

gen employ = 0
replace employ = 1 if indic==1 & inlist(irwrkstat18,1,2,3,4)
replace employ = 1 if indic==2 & inlist(irwrkstat18,1)

/*Create the subpopulation variable*/
generate subpop = 1 if catag18 == 1 & employ == 1
svyset verep [pweight=analwt2], strata(vestr) dof(750)
svy, subpop(subpop): mean cigmon, over(indic)
test c.cigmon@1.indic = c.cigmon@2.indic

/*Since subsetting to employ=1, this is testing all persons 18+
vs. employed persons 18+ for past month cigarette use*/
/* employ is defined earlier in this exhibit and catag18=1 for
persons 18 or older and 2 otherwise*/

```

**Exhibit A.3.11 Stata Code (Test for Independence Based on a Log-Linear Model)**

This exhibit 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 (variable CIGMON) is associated with employment status (variable IRWRKSTAT18). DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```

use using ".\\dataname_single.dta", clear
/*Ensure all variables are lower case*/
rename *, lower

/*Need to subset to just 4 levels of irwrkstat18*/
generate subpop = 1 if inlist(irwrkstat18,1,2,3,4)
/*four level employment status variable*/

svyset verep [pw=analwt2], strata(vestr) dof(750)
svy, subpop(subpop): tab cigmon irwrkstat18, llwald noadjust

/*This will give you both the adjusted and non-adjusted Wald F,
the non-adjusted test statistic will match SUDAAN*/

```

**Exhibit A.3.12 Stata Code (Pairwise Testing)**

Once an association is determined by a significant Wald *F* test computed in [Exhibit A.3.11](#), then pairwise testing as shown below can be done to determine individual significant tests across levels of an independent variable (e.g., employment status). DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```

use using ".\\dataname_single.dta", clear
/*Ensure all variables are lower case*/

```

### Exhibit A.3.12 Stata Code (Pairwise Testing) (continued)

```

rename *, lower

/*Need to subset to just 4 levels of irwrkstat18*/
generate subpop = 1 if inlist(irwrkstat18,1,2,3,4)
/*four level employment status variable*/

svyset verep [pw=analwt2], strata(vestr) dof(750)

/*Estimated means of past month cigarette use by employment
status*/
svy, subpop(subpop): mean cigmon, over(irwrkstat18)
matrix Me = e(b)

local max=4 /*number of irwrkstat18 categories*/
matrix output = J(6,7,.) /*empty matrix to store results; the
number of rows should match the number of contrasts needed*/

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 c.cigmon@`j'.irwrkstat18 =
            c.cigmon@`stop'.irwrkstat18, 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
}
/*Estimated Totals*/
svy: total cigmon, over(irwrkstat18)

matrix M = e(b) /*Store total estimates in matrix M*/
local max=4 /*number of categories*/

local counter1 = `max' - 1
local counter2 = `max' - 1
local contrast = 0

forvalues i=1/`counter1' {
    local stop = `max' - `i' + 1
    forvalues j=1/`counter2' {

```

**Exhibit A.3.12 Stata Code (Pairwise Testing) (continued)**

```

    local contrast = `contrast' + 1
    test c.cigmon@`j'.irwrkstat18 = c.cigmon@`stop'.irwrkstat18,
    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 ///
    settotal mean_pval
matrix list output

```

**Exhibit A.3.13 Stata Code (Test of Linear Trends Across Years)**

This exhibit displays code that tests for a linear trend in past month alcohol use (variable ALCMON) by sex across 4 years (variable YEAR) to evaluate change over time. This example uses a nonparametric method of calculation, similar to SUDAAN [Exhibit A.2.13](#). Users can refer to [Table A.3](#) to assist in developing an appropriate contrast to be used with the “coeff” function. DATANAME is a dataset that has 4 years of data with a year variable defined for each year on the dataset.

```

use using ".\\dataname.dta", clear
rename *, lower
svyset verep [pw=analwt2], strata(vestr) dof(750)

/*Estimated Means*/
svy: mean alcmon, over(year irsex)
matrix Me = e(b)
matrix coeff = (-3, -1, 1, 3) /*Define coeff as the coefficients
according to the # of years, see Table A.3. Note that the coefficients
must be separated by commas*/

local max=4*2 /* Max is the Total Number of Years multiplied by the
Total levels of IRSEX*/
local counter1 = 2

generate pmean=.
generate mean=.
generate semean=.

forvalues i=1/ `counter1' {
    local stop = `max' / `counter1'
    local test
    local mean
    forvalues j=1/`stop' {
        local sub = `i' + `counter1'*(`j'-1)
        local co = coeff[1,`j']
    }
}

```

### Exhibit A.3.13 Stata Code (Test of Linear Trends Across Years) (continued)

```

    local test = "`test' (`co')*c.alcmon@`j'.year#`i'.irsex"
    local mean = "`mean' `co'*Me[1,`sub']"
    if (`j' < `stop') {

local stop = `max' / `counter1'
local test
local total
forvalues j=1/`stop' {
    local sub = `i' + `counter1'*(`j'-1)
    local co = coeff[1, `j']
    local test = "`test' (`co')*c.alcmon@`j'.year#`i'.irsex"
    local total = "`total' `co'*M[1, `sub']"
    if (`j' < `stop') {
        local test = "`test' + "
        local total = "`total' + "
    }
}
test `test' = 0, nosvyadjust matvlc(test`counter1')
replace setotal= sqrt((test`counter1'[1,1])) if irsex==`i'
replace total=`total' if irsex==`i'
}

keep irsex mean semean total setotal pmean
duplicates drop irsex mean semean total setotal pmean, force

/*Keep one record per contrast*/
drop if total ==.
format pmean %-15.10f
format total %-12.0fc
format setotal %-12.0fc

/*Output the results*/
list irsex mean semean total setotal pmean

```

## A.4 SAS Exhibits

The SAS exhibits in this section provide guidance on how to use various programming options to produce estimates for NSDUH using the statistical procedures documented within this report.

### Exhibit A.4.1 SAS SURVEYMEANS Procedure (Estimate Generation: Single Year and Pooled Years of Data)

This exhibit demonstrates how to compute various types of estimates for past month alcohol use by year and sex for single- or combined-year (pooled) data using the SAS SURVEYMEANS procedure. SAS SURVEY procedures do not require either data sorting or specifying the degrees of freedom before analysis of complex survey data. The default variance estimation is the Taylor series linearization method.

Code is included to compute the prevalence estimate, SE of the mean, weighted sample size, unweighted sample size, weighted total, and SE of the totals. 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 (i.e., domains forced to match their respective U.S. Census Bureau and ACS population estimates through the weight calibration process). For additional information on SEs, see Section A.1.5. For more information on how to create a pooled weight to use when producing annual averages of combined years of data, see Chapter 3. The input dataset would need to include at least 2 years of NSDUH data with a YEAR variable defined for each year on the dataset.

```
TITLE "ESTIMATES OF PAST MONTH ALCOHOL BY YEAR AND SEX";
PROC SURVEYMEANS DATA=DATANAME SUMWGT NOBS MEAN SUM;
STRATA VESTR; /*Nesting variable - strata*/
CLUSTER VEREP; /*Nesting variable - PSU*/
WEIGHT ANALWT2; /*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.*/
VAR ALCMON; /*Past month alcohol analysis variable*/
DOMAIN YEAR*IRSEX; /*Sex by year*/
    /*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*/
    /*Sex variable, where male=1 & female=2*/
ODS OUTPUT DOMAIN=OUT.SASFILE;
RUN;
```

### Exhibit A.4.2 SAS Code Based on SAS Output (Calculation of Standard Error of Totals for Fixed Domains)

This exhibit estimates the SE of the total for fixed domains by using the data produced in [Exhibit A.4.1](#). Because sex is a fixed domain, the SE of the totals would not be taken directly from the examples in [Exhibit A.4.1](#) but rather would be computed using the alternative SE estimation method as shown below. 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:

**Exhibit A.4.2 SAS Code Based on SAS Output (Calculation of Standard Error of Totals for Fixed Domains) (continued)**

```
DATA SASEST;
SET OUT.SASFILE; /*input the output file from above SAS procedure
in Exhibit A.4.1 */

IF IRSEX IN (1,2) THEN SETOTAL=SUMWGT*STDERR;

RUN;
```

**Exhibit A.4.3 SAS Code Based on SAS Output (Implementation of Suppression Rule)**

This exhibit applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1 (commented out in the exhibit) using the data produced in [Exhibit A.4.1](#). Starting with the 2024 NSDUH, the suppression rule was simplified, and the code for the new rules described in [Table 9.1](#) is shown in this exhibit.

```
DATA SASEST;
SET OUT.SASFILE;

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

/*SUPPRESSION RULE FOR PREVALENCE ESTIMATES*/
IF (STDERR=0) OR (RSELNP GT 0.175) OR (N <50) 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;
```

**Exhibit A.4.4 SAS Code (Tests of Differences)**

This exhibit performs significance testing between comparable years of NSDUH data using the data produced by the example in [Exhibit A.4.1](#). The input dataset would need to include at least 2 years of NSDUH data, but SAS does not require the data to be sorted by nesting variables. Creation of the YEAR variable is omitted from the example. 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.

```
TITLE "TESTS OF DIFFERENCES BETWEEN YEAR1 AND YEAR2 ESTIMATES OF PAST
MONTH ALCOHOL BY SEX";
PROC SURVEYREG DATA=DATANAME;
```

**Exhibit A.4.4 SAS Code (Tests of Differences) (continued)**

```

CLUSTER VEREP;
STRATA VESTR;
WEIGHT ANALWT2;
DOMAIN IRSEX;
CLASS YEAR;
MODEL ALCMON = YEAR /NOINT VADJUST=NONE SOLUTION COVB;
/* option NOINT omits the intercept from the model; option
VADJUST=NONE specifies that no variance adjustment is used;
option SOLUTION displays the parameter estimates; option
COVB displays the estimated covariance matrix of the
estimated regression estimates. */
LSMEANS YEAR / DIFF ;
ODS OUTPUT DIFFS = OUT.SASTESTS /*output the estimates of
difference*/
COVB = OUT.COVB; /*output the variance covariance matrix*/
RUN;
/*Note: to compare other years, more years could be added to the
dataset and a WHERE clause could be added to restrict the test to the
specified years where X = year1 and Y = year2: WHERE YEAR IN (X, Y);
*/

```

**Exhibit A.4.5 SAS Code (Calculation of the *P* Value for the Test of Differences between Totals for Nonfixed Domains)**

Note that in SAS, the SE of the difference between totals cannot be produced. The current SAS procedure only provides SE of difference between means.

**Exhibit A.4.6 SAS Code (Covariance Matrix, Calculations of Variances, and Calculation of the *P* Value for the Test of Differences between Totals for Fixed Domains)**

To calculate the *p* value for the test of differences between totals for fixed domains, three SAS datasets, one containing the covariances (SASCOV) and two containing the variances (EST1 and EST2), are then merged with the output dataset (SASTESTS) from the procedure that generated the tests of differences in [Exhibit A.4.4](#).

```

/*Option used to allow trailing blanks in variable names*/
OPTIONS VALIDVARNAME=ANY;

TITLE "CALCULATE THE P-VALUE FOR THE TEST OF DIFFERENCE BETWEEN FIXED
DOMAINS";

/*Bring in covariance dataset from Exhibit A.4.4*/
TITLE "COVARIANCE MATRIX";
DATA OUT.SASCOV (KEEP=IRSEX COV);
SET OUT.COVB;

    IF Parameter = 'year2324 1' THEN COV = year2324_2;

/* Assign IRSEX based on DOMAIN label */
    IF DOMAIN = "Imputation-Revised Sex of Respondent=1" THEN
        IRSEX = 1;

```

**Exhibit A.4.6 SAS Code (Covariance Matrix, Calculations of Variances, and Calculation of the *P* Value for the Test of Differences between Totals for Fixed Domains) (continued)**

```

ELSE IF DOMAIN = "Imputation-Revised Sex of Respondent=2" THEN
  IRSEX = 2;
  ELSE IRSEX = 0;

/* Output only relevant rows */
IF COV NE . THEN OUTPUT;
RUN;
```

The variances of the means are estimated in the following separate data steps and output into datasets EST1 and EST2. The variance is the square of the SE of the mean. The SEs of the means were output in the original procedure that generated the estimates in [Exhibit A.4.1](#).

```

/*Bring in dataset from Exhibit A.4.1 and Calculate the Variances*/
TITLE "TESTS OF DIFFERENCES BETWEEN TOTALS";
DATA EST1 (KEEP=N1 WSUM1 VAR1 TOTAL1 VAR_TOTAL1 IRSEX);
SET OUT.SASFILE;
  WHERE YEAR=1;
  N1=N;
  WSUM1=SUMWGT;
  VAR1=STDERR**2;
  TOTAL1=SUM;
  VAR_TOTAL1=STDDEV**2;
  IF IRSEX=. THEN IRSEX=0;
RUN;

DATA EST2 (KEEP=N2 WSUM2 VAR2 TOTAL2 VAR_TOTAL2 IRSEX);
SET OUT.SASFILE;
  WHERE YEAR=2;
  N2=N;
  WSUM2=SUMWGT;
  VAR2=STDERR**2;
  TOTAL2=SUM;
  VAR_TOTAL2=STDDEV**2;
  IF IRSEX=. THEN IRSEX=0;
RUN;

/*Bring in estimate of differences dataset from Exhibit A.4.4*/
DATA OUT.SASTESTS_CLEAN;
SET OUT.SASTESTS;
  /*Confirm variable label as labels can change from year to year*/
  IF DOMAIN = "Imputation-Revised Sex of Respondent=1" THEN
    IRSEX=1;
  ELSE IF DOMAIN = "Imputation-Revised Sex of Respondent=2" THEN
    IRSEX=2;
  ELSE IRSEX=0;
RUN;
```

#### Exhibit A.4.6 SAS Code (Covariance Matrix, Calculations of Variances, and Calculation of the $P$ Value for the Test of Differences between Totals for Fixed Domains) (continued)

With the proper statistics contained in one dataset (data P\_VALUE), the corresponding  $p$  value for the tests of differences between fixed domain totals can be produced using the SAS PROBT function and calculated  $t$  test statistic as shown below.

```
/*Create P-value*/
DATA OUT.P_VALUE;
MERGE EST1_EST2 OUT.SASTESTS_CLEAN OUT.SASCOV;
  BY IRSEX;
  TOTAL=TOTAL1-TOTAL2;
  SE_DIFF = SQRT(WSUM1**2*VAR1 + WSUM2**2*VAR2 -
  2*WSUM1*WSUM2*COV);
  PVALT = 2 * (1 - PROBT (ABS (TOTAL / SE_DIFF), 750));
RUN;
```

#### Exhibit A.4.7 SAS Code (Recoding a Variable, Estimate Generation with (1) Missing Values and (2) Using Subpopulation)

This exhibit shows two methods for handling unused values in variable recodes in estimate generation using a single year of NSDUH data. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```
/*Method 1, recoding unused values to missing*/
TITLE "PRODUCE ESTIMATES WHERE THE VARIABLE HAS MISSING VALUES";
DATA RECODES;
SET DATANAME_SINGLE;
  IF IRMJAGE=991 THEN IRMJAGE_R=.;
  ELSE IRMJAGE_R=IRMJAGE;
RUN;

PROC SURVEYMEANS DATA=RECODES;
  CLUSTER VEREP;
  STRATA VESTR;
  WEIGHT ANALWT2;
  VAR IRMJAGE_R;
RUN;

PROC SURVEYMEANS DATA=RECODES;
  CLUSTER VEREP;
  STRATA VESTR;
  WEIGHT ANALWT2;
  DOMAIN IRSEX; /*Estimates by Sex*/
  VAR IRMJAGE_R;
RUN;

/*Method 2, using subpopulation to omit the unused values*/
PROC SURVEYMEANS DATA=DATANAME_SINGLE;
  WHERE MRJFLAG=1;
  CLUSTER VEREP;
```

**Exhibit A.4.7 SAS Code (Recoding a Variable, Estimate Generation with (1) Missing Values and (2) Using Subpopulation) (continued)**

```

STRATA VESTR;
WEIGHT ANALWT2;
VAR IRMJAGE;

RUN;

PROC SURVEYMEANS DATA=DATANAME_SINGLE;
WHERE MRJFLAG=1;
CLUSTER VEREP;
STRATA VESTR;
WEIGHT ANALWT2;
DOMAIN IRSEX; /*Estimates by Sex*/
VAR IRMJAGE;

RUN;

```

**Exhibit A.4.8 SAS Code (Calculates a Confidence Interval for Alcohol Drinker Prevalence and Estimated Totals Produced in [Exhibit A.4.1](#))**

This exhibit computes a 95 percent CI. 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 CI of the mean, as shown below. 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.

```

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. Code provided below uses output from [Exhibit A.4.1](#) to create the 95 percent CIs for means and totals using manual coding steps for calculations in SAS. The CI 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.

```

TITLE "CALCULATE CONFIDENCE INTERVALS";
DATA OUT.CI;
SET OUT.SASFILE; /*output data from Exhibit A.4.1*/
T_QNTILE=TINV(0.975,750); /*define t-statistic*/
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));
/*PLOWER AND PUPPER ARE THE 95% CIS ASSOCIATED WITH MEAN*/

TLOWER=SUMWGT*PLOWER;
TUPPER=SUMWGT*PUPPER;
/*TLOWER AND TUPPER ARE THE 95% CIS ASSOCIATED WITH TOTAL*/

RUN;

```

**Exhibit A.4.9 SAS (Estimate Generation for Categorical Variable; i.e., Number of Days Used Substance in the Past Month among Past Month Users)**

This exhibit shows how to compute estimates corresponding to levels of a categorical variable. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```
TITLE "CALCULATE PERCENTAGES AND ASSOCIATED SEs";
PROC SURVEYFREQ DATA=DATANAME_SINGLE;
  WHERE MRJMON=1;
  CLUSTER VEREP;
  STRATA VESTR;
  WEIGHT ANALWT2;
  TABLE MRJMDAYS;
RUN;
```

**Exhibit A.4.10 SAS Code (Statistical Tests of Differences between Two Groups when the Two Groups Overlap)**

This exhibit shows how to test among overlapping domains (i.e., scenarios where some cases are in both domains are being compared). A stacked dataset is created first that includes two records for each respondent in the overlap needed for analysis. The stacked data are then used to compute the test of differences among the overlapping domains. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```
TITLE "PERFORM TEST OF DIFFERENCE BETWEEN TWO OVERLAPPING GROUPS";
DATA STACKED;
  SET DATANAME_SINGLE(IN=A) DATANAME_SINGLE(IN=B); /*reading in
  data twice*/
  /*create an indicator variable for the stacked data, this will be
  used in PROC SURVEYREG
  When indic=1, employ=1 represents the full population
  When indic=2, employ=1 represents those employed full time*/
  IF A THEN DO;
    INDIC=1;
    IF IRWRKSTAT18 IN (1,2,3,4) THEN EMPLOY=1;
    /*IRWRKSTAT18 is a four-level employment variable for adults,
    where level 1 is those employed full time, 2 is those employed
    part time, 3 are those unemployed, and 4 are all other adults.
    Respondents aged 12 to 17 are coded as level 99*/
    ELSE EMPLOY=0;
  END;
  ELSE IF B THEN DO;
    INDIC=2;
    IF IRWRKSTAT18=1 THEN EMPLOY=1;
    ELSE EMPLOY=0;
  END;

RUN;

PROC SURVEYREG DATA=STACKED;
  WHERE CATAG18=1 & EMPLOY=1;
  CLUSTER VEREP;
```

**Exhibit A.4.10 SAS Code (Statistical Tests of Differences between Two Groups when the Two Groups Overlap) (continued)**

```

STRATA VESTR;
WEIGHT ANALWT2;
CLASS INDIC;
MODEL CIGMON=INDIC/NOINT VADJUST=NONE SOLUTION COVB;
LSMEANS INDIC/DIFF;
RUN;

```

**Exhibit A.4.11 SAS Code (Tests of the Independence of the Prevalence Variable and Subgroup Variable)**

This exhibit shows how to create a test of independence for a subpopulation defined by three or more levels of a categorical variable. This code calculates both a Rao-Scott chi-square test (CHISQ) and the log-linear chi-square Wald  $F$  test (WLLCHISQ) to indicate whether cigarette use is associated with employment status. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```

TITLE "PERFORM TEST OF INDEPENDENCE OF THE PREVALENCE VARIABLE AND
SUBGROUP VARIABLE";
PROC SURVEYFREQ DATA=DATANAME_SINGLE;
  WHERE IRWRKSTAT18 IN (1,2,3,4);
  CLUSTER VEREP;
  STRATA VESTR;
  WEIGHT ANALWT2;
  TABLE IRWRKSTAT18*CIGMON / COL ROW CHISQ WLLCHISQ; /*option COL
  displays column percentages; option ROW displays row percentages;
  option CHISQ requests Rao-Scott chi-square test; option WLLCHISQ
  requests Wald log-linear chi-square test. */
  ODS OUTPUT WLLCHISQ=OUT.SAS_CHI;
RUN;

```

**Exhibit A.4.12 SAS Code (Pairwise Testing)**

Once an association is determined by a significant chi-square test computed in [Exhibit A.4.11](#), then pairwise testing as shown below can be done to determine individual significant tests across levels of employment status. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```

TITLE "PERFORM PAIRWISE TESTS FOR EACH SUBGROUP VARIABLE";
PROC SURVEYREG DATA=DATANAME_SINGLE;
  WHERE IRWRKSTAT18 IN (1,2,3,4);
  CLUSTER VEREP;
  STRATA VESTR;
  WEIGHT ANALWT2;
  CLASS IRWRKSTAT18;
  MODEL CIGMON=IRWRKSTAT18/NOINT VADJUST=NONE SOLUTION COVB;
  LSMEANS IRWRKSTAT18/DIFF;
RUN;

```

### Exhibit A.4.13 SAS Code (Test of Linear Trends Across Years)

This exhibit displays code that tests for a linear trend in past month alcohol use (variable ALCMON) by sex across 4 years (variable YEAR) to evaluate change over time. Unlike the nonparametric approach used in SUDAAN ([Exhibit A.2.13](#)), this example uses the SAS procedure SURVEYLOGISTIC, which is a model-based method. This method allows for testing changes over time while controlling for additional covariates. Users should note the following differences between the procedure used in this example and the nonparametric procedure used in SUDAAN [Exhibit A.2.13](#):

- PROC SURVEYLOGISTIC does not automatically provide estimates for the contrast total or SEs of the contrast total.
- A DOMAIN statement is required to perform the analysis for levels of a stratification variable (IRSEX in this example).
- The variable YEAR is treated as a continuous predictor instead of a categorical variable.
- The COVB option is included to generate the covariance matrix of the parameter estimates.
- The SURVEYLOGISTIC output includes the odds ratio and its CI.

The calculated test statistic and  $p$  value for the model-based method will likely differ from the nonparametric method, but the overall interpretation of the results should be consistent. DATANAME is a dataset that has 4 years of data with a year variable defined for each year on the dataset.

```
PROC SURVEYLOGISTIC DATA=DATANAME;  
DOMAIN IRSEX;  
CLUSTER VEREP;  
STRATA VESTR;  
WEIGHT ANALWT2;  
MODEL ALCMON(EVENT='1')=YEAR / COVB;  
RUN;
```

## A.5 R Code Exhibits

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

### Exhibit A.5.1 R Code: svytotal and svymean (Estimate Generation: Single Year and Pooled Years of Data)

This exhibit displays code that calculates the unweighted count, prevalence (mean), totals, and design effect for past month alcohol use (variable ALCMON) by sex for a single year and pooled years of data. When analyzing pooled years of data, a modified analysis weight must be used. Creation of the YEAR variable is omitted from this example. The input dataset for this example is a pooled 2-year dataset. Appropriate estimates are generated by specifying the proper complex survey design object. For a single year analysis, the "design object" is used. For a 2-year analysis, the "combined\_design" object is used.

```
# set directory, install packages (only need to run once)
# note that R is case sensitive
#####
#install.packages("haven")
#install.packages("survey")
#install.packages("dplyr")
#install.packages("multcomp")

library(haven)
library(survey)
library(dplyr)
library(multcomp)

# read NSDUH SAS files: YEAR1, YEAR2, 2-year data with selected
# variables
#####
# use haven to read NSDUH SAS dataset
# use DATANAME as generic dataframe object throughout for 2-year file
# YEAR variable, where YEAR1=1 & YEAR2=2. Alternatively, the year
# variable could identify the combined years of data, i.e., YEAR1 and
# YEAR2 = 1 and YEAR3 and YEAR4 = 2

DATANAME<-read_sas("SAS DATASET PATH AND NAME",
col_select=c(QUESTID, VESTR, VEREP, ANALWT2, YEAR, IRSEX, ALCMON, MRJMDAYS, MR
JMON, CIGMON, IRWRKSTAT18, CATAG18, IRMJAGE, MRJFLAG))

#convert variable names to lower case
names(DATANAME)<-tolower(names(DATANAME))

# Create updated weight for a pooled 2 year data analysis
DATANAME$analwt2_2yr <- DATANAME$analwt2/2

#####
```

**Exhibit A.5.1 R Code: svytotal and svymean (Estimate Generation: Single Year and Pooled Years of Data) (continued)**

```

Create_design <- function(dataname,weight_var) {
svydesign(
  id = ~ verep ,
  strata = ~ vestr ,
  data =dataname ,
  weights = as.formula(paste("~",weight_var)),
  nest = TRUE
) %>%
update(
  one = 1 ,
  yearfactor =
  factor(
    year ,
    levels = 1:2 ,
    labels = c( "YEAR1" , "YEAR2" ) ) ,
  irsex =
  factor(
    irsex ,
    levels = 1:2 ,
    labels = c( "male" , "female")),
  mrjmdays =
  factor(
    mrjmdays,
    levels = 1:5 ,
    labels = c("1=1-2 days","2=3-5 days","3=6-19 days","4=20+
days","5=did not use in the past month")),

yearcombined=ifelse(year %in% c("YEAR1", "YEAR2"), 1, 0),
year0=ifelse(year==1, 1, 0),
year1=ifelse(year==2, 1, 0),
sexmale=ifelse(irsex=="male", 1, 0),

sexfemale=ifelse(irsex=="female", 1, 0) )
}
#Create survey designs for both one and two year data
design(DATANAME, "analwt2")
Combined_design <- Create_design(DATANAME, "analwt2_2yr")

# degrees of freedom
degf(design)

# sample domain N
##pooled two year
sum(weights(design , "sampling" ) != 0 )

##each year (1=YEAR1; 2=YEAR2)
svyby( ~ one , ~ year , design , unwt.d.count )

```

### Exhibit A.5.1 R Code: svytotal and svymean (Estimate Generation: Single Year and Pooled Years of Data) (continued)

```
##sex in pooled two years
svyby( ~ one , ~ irsex , combined_design , unwtd.count )
##sex by each year
svyby( ~ one , ~ year+irsex , design , unwtd.count )

# weighted sample domain N
##each year (1=YEAR1; 2=YEAR2)
svytotal(~one, combined_design) %>% round
#Pooled two years
svyby(~one, ~year, design, FUN=svytotal) %>% round
## sex in pooled two year
svyby( ~ one , ~ irsex , combined_design , FUN=svytotal)
## by sex and year
svyby( ~ one , ~ year+irsex , design , FUN=svytotal )

# proportion estimate: past month Alcohol use
##pooled two years
svymean(~alcmon, combined_design, deff = "replace") %>% round(2)

##each year
svyby(~alcmon, ~year, design, svymean, deff = "replace" )
##sex in pooled two years
svyby(~alcmon, ~irsex, combined_design, svymean, deff = "replace" )
##sex by each year
svyby(~alcmon, ~year+irsex, design, svymean, deff = "replace" )

# count estimate: past month Alcohol drinker number total
#pooled two years
svytotal(~alcmon, combined_design)
# by year
svyby(~alcmon, ~year, design, svytotal )
#by sex in pooled two years
svyby(~alcmon, ~irsex, combined_design, svytotal )
#by sex year
svyby(~alcmon, ~year+irsex, design, svytotal )
```

### Exhibit A.5.2 R Code (Calculation of Standard Error of Totals for Fixed Domains)

This exhibit computes the SE of the total for fixed domains using the data produced by the example in [Exhibit A.5.1](#). Because sex is a fixed domain, the SE of the totals would not be taken directly from the examples in [Exhibit A.5.1](#) but rather would be computed using the alternative SE estimation method as shown below.

```
# sex in NSDUH is a fixed domain. Accordingly, for count
# estimate of past month Alcohol drinkers, corrected SE is computed
# here.

# compute the corrected SE by sex and year here.
# weighted sample domain N by sex year
wdomain=svyby(~one, ~year+irsex, design, svytotal )
```

### Exhibit A.5.2 R Code (Calculation of Standard Error of Totals for Fixed Domains) (continued)

```
# 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
#end of computation for corrected SE

# Repeat for combined sex by year
wdomaintot=svyby(~one, ~year, design, svytot )
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
#end of computation for corrected SE
```

### Exhibit A.5.3 R Code (Implementation of Suppression Rule)

This exhibit applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1 (commented out in the exhibit) using the data produced by the example in [Exhibit A.5.1](#). Starting with the 2024 NSDUH, the suppression rule was simplified, and the code for the new rules described in [Table 9.1](#) is shown in this exhibit.

```
# Here, we focus on suppression rule for proportion estimates.
# See Table 9.1 for other estimates
## proportion estimate of Alcohol use by sex and year
prop=svyby(~alcmon, ~year+irsex, design, svymean, deff = "replace");
prop
ndomain=svyby( ~ one , ~ year+irsex , design , unwtd.count ); ndomain

# sample domain N by sex and year
##combine together
prop=cbind(prop, subset(ndomain, select=c(counts))); prop

## Compute relative standard error (RSE) of national log P
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))

#Suppression rule for proportion estimates: if suprurle=1 then
suppress;
#do not if suprurle=NA
prop$suprurle=ifelse(prop$alcmon ==0 | prop$RSELNP > 0.175
| prop$count <50, 1, NA); prop

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

**Exhibit A.5.4 R Code (Tests of Differences)**

This exhibit 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.

```
#We make significance test of difference in proportion estimates of
# alcohol drinkers between 2 years, separately for each of the
# following domains:
# 1) total pop; 2) male only; 3) female only
## sample count: 1=YEAR1, 2=YEAR2
count(DATANAME, year)
## sample count: 1=male, 2=female
count(DATANAME, irsex)

# 1) total pop
## proportions by year
svyby(~alcmon, ~year, design, svymean )

# sig. testing: Alcohol drinker proportion difference=year2 - year1
svyttest(alcmon~year, design)

# 2) male only
## proportions within male
svyby(~alcmon, ~year, subset(design , irsex == "male" ), svymean )

# sig. testing: Alcohol drinker proportion difference: year1 vs year2
# within male
svyttest( alcmon ~ year , subset(design , irsex == "male" ) )

# 3) female only
## proportions within female

svyby(~alcmon, ~year, subset(design , irsex == "female" ), svymean )
# sig. testing: Alcohol drinker proportion difference: year1 vs year2
# within female
svyttest( alcmon ~ year , subset(design , irsex == "female" ) )
```

**Exhibit A.5.5 R Code (Calculation of the *P* Value for the Test of Differences between Estimated Number Totals for Nonfixed Domains)**

With the statistics and the correct degrees of freedom (750 in this example), the *p* value (PvalueT) for the test of differences between totals for nonfixed domains can be calculated as shown below. The R function pt returns the probability from a *t*-distribution.

```
# Sig. testing of difference in past month Alcohol drinker number
# between two years, separately in each of the following domains:
# 1) total pop; 2) male only; 3) female only

# year and sex are fixed domains, but here we pretend that they are
# non-fixed domains and compute p values.
```

**Exhibit A.5.5 R Code (Calculation of the  $P$  Value for the Test of Differences between Estimated Number Totals for Nonfixed Domains) (continued)**

```

# 1) among total pop
# Difference in alcohol drinker numbers between two years
# (YEAR1 versus YEAR2)

##estimated number of alcohol drinker
total=svytotal(~I(alcmon*year0)+I(alcmon*year1), design); total
contrast=svycontrast(total, list(diff=c(1,-1))); contrast

#calculation of p value for test of differences between totals for
#nonfixed domains
pvalueT=2*(1-pt(abs(coef(contrast)/SE(contrast)),750)); pvalueT

# 2) among male
#Difference in alcohol drinker numbers between two years among males
#(YEAR1.male versus YEAR2.male)

#estimated number of alcohol drinker
total=svytotal(~I(alcmon*year0)+I(alcmon*year1), design=subset(design,
irsex=="male")); total
contrast=svycontrast(total, list(diff=c(1,-1))); contrast
#calculation of p value for test of differences between totals for
#nonfixed domains
pvalueT=2*(1-pt(abs(coef(contrast)/SE(contrast)),750)); pvalueT

# 3) among female
#Difference in alcohol drinker numbers between two years among females
#(YEAR1.female versus YEAR2.female)

##estimated number of alcohol drinker
total=svytotal(~(alcmon*year0)+I(alcmon*year1), design=subset(design,
irsex=="female")); total
contrast=svycontrast(total, list(diff=c(1,-1))); contrast

#calculation of p value for test of differences between totals for
#nonfixed domains
pvalueT=2*(1-pt(abs(coef(contrast)/SE(contrast)),750)); pvalueT

```

**Exhibit A.5.6 R Code (Covariance Matrix, Calculations of Variances, and Calculation of the  $P$  Value for the Test of Differences between Totals for Fixed Domains)**

This exhibit shows how to calculate the  $p$  value for the test of differences between totals for fixed domains using R. Each individual piece is computed that is needed to calculate the  $t$ -statistic. First, the covariances are computed separately for the total population, males, and females.

```

# We do the same sig. testing of difference as we did in
# Exhibit A.5.4. The difference is that here we do so correctly
# because sex and year are fixed domains.

```

### Exhibit A.5.6 R Code (Covariance Matrix, Calculations of Variances, and Calculation of the *P* Value for the Test of Differences between Totals for Fixed Domains) (continued)

```
# Here, we make sig. testing of two estimates of alcohol drinkers
# (YEAR1 vs YEAR2) by computing the correct p value
# we do so separately in each of the three population groups:
# 1) total pop; 2) male; 3) female
```

```
#pulling the relevant covariance of alcohol drinker proportion
# estimates #for two years the covariance matrix of proportion
# estimates between 2 years
```

```
## 1) total pop
prop1=svyglm(alcmon~yearfactor, design); vcov1=vcov(prop1)
```

```
## 2) male only
prop2=svyglm(alcmon~yearfactor, subset(design, irsex=="male"));
vcov2=vcov(prop2)
```

```
# 3) female only
prop3=svyglm(alcmon~yearfactor, subset(design, irsex=="female"));
vcov3=vcov(prop3)
cov1=vcov1[1,1]+vcov1[2,1] # covariance (1) total pop
cov2=vcov2[1,1]+vcov2[2,1] # covariance (2) male
cov3=vcov3[1,1]+vcov3[2,1] # covariance (3) female
```

Then the SEs are calculated separately for the total population, males, and females.

```
# calculate the SE of alcohol drinker proportion estimates by year
## 1) total Pop
se1=svyby(~alcmon, ~year, design, svymean );SE(se1)
## 2) male
se2=svyby(~alcmon, ~year, subset(design , irsex == "male" ), svymean
); SE(se2)
## 3) female
se3=svyby(~alcmon, ~year, subset(design , irsex == "female" ), svymean
); SE(se3)
```

Then the difference in the totals is calculated for the total population, males, and females.

```
# difference in estimated alcohol drinker number totals
##estimated number of alcohol drinker: 1) total
total1=svytotal(~I(alcmon*year0)+I(alcmon*year1), design); total1
## difference 1) among total pop
contrast1=svycontrast(total1, list(diff=c(1,-1))); contrast1

##estimated number of alcohol drinker: 2) male
total2=svytotal(~I(alcmon*year0)+I(alcmon*year1),
design=subset(design, irsex=="male")); total2
# difference in male (YEAR1.male versus YEAR2.male)
contrast2=svycontrast(total2, list(diff=c(1,-1))); contrast2
```

### Exhibit A.5.6 R Code (Covariance Matrix, Calculations of Variances, and Calculation of the *P* Value for the Test of Differences between Totals for Fixed Domains) (continued)

```
#estimated number of alcohol drinker: 3) female
total3=svytotal(~I(alcmon*year0)+I(alcmon*year1),
design=subset(design, irsex=="female")); total3
# difference in female (YEAR1.female versus YEAR2.female)
contrast3=svycontrast(total3, list(diff=c(1,-1))); contrast3
```

Then the weighted sample size is calculated for the total population, males, and females.

```
# weighted sample N by year
# 1) total pop
wdomain1=svyby(~one, ~year, design, svytotal ); coef(wdomain1)
# 2) male
wdomain2=svyby(~one, ~year, subset(design , irsex == "male" ),
svytotal ); coef(wdomain2)
# 3) female
wdomain3=svyby(~one, ~year, subset(design , irsex == "female" ),
svytotal ); coef(wdomain3)
```

Using all the measures computed above, the corresponding *p* value for the tests of differences between fixed domain totals can be produced for the total population, males, and females.

```
#Calculate p value for three comparisons (all, male, female)
#calculation of p value for test of differences between estimated
#alcohol drinker user totals for Fixed domains

# 1) total pop
pvalueT1=2*(1-
pt(abs(contrast1[1]/sqrt(coef(wdomain1)[1]^2*SE(se1)[1]^2+coef(wdomain
1)[2]^2*SE(se1)[2]^2
-2*coef(wdomain1)[1]*coef(wdomain1)[2]*cov1)),750));
pvalueT1

# 2) male
pvalueT2=2*(1-
pt(abs(contrast2[1]/sqrt(coef(wdomain2)[1]^2*SE(se2)[1]^2+coef(wdomain
2)[2]^2*SE(se2)[2]^2
-
2*coef(wdomain2)[1]*coef(wdomain2)[2]*cov2)),750));pvalueT2

# 3) female
pvalueT3=2*(1-
pt(abs(contrast3[1]/sqrt(coef(wdomain3)[1]^2*SE(se3)[1]^2+coef(wdomain
3)[2]^2*SE(se3)[2]^2
-
2*coef(wdomain3)[1]*coef(wdomain3)[2]*cov3)),750));pvalueT3
```

**Exhibit A.5.7 R Code (Estimate Generation with (1) Missing Values and (2) Using Subpopulation)**

This exhibit shows two methods for handling unused values in variable recodes in estimate generation using a single year of NSDUH data. DATANAME\_SINGLE is a data frame that has been subset to a single year of data.

```
# Marijuana Age of First Use recoded analysis variable is used
# Data management and create survey design

##recode as
DATANAME_SINGLE$irmjage_r=ifelse(DATANAME_SINGLE$irmjage==991,
NA,DATANAME_SINGLE$irmjage)

design.R <-
  svydesign(
    id = ~ verrep ,
    strata = ~ vestr ,
    data = DATANAME_SINGLE ,
    weights = ~ analwt2 ,
    nest = TRUE
  )

design.R <-
  update(design.R,
    one = 1 ,

    yearfactor =
      factor(
        year ,
        levels = 1:2 ,
        labels = c( "YEAR1" , "YEAR2" ) ) ,
    irsex =
      factor(
        irsex ,
        levels = 1:2 ,
        labels = c( "male" , "female" ))
  )

#Mean Marijuana Age of First Use and its SE
#Method1: recoding those who never used Marijuana as missing
mean1=svyby( ~ irmjage_r , ~ irsex , design.R , svymean , na.rm = TRUE
); mean1
#Method2: using subpopulation to omit those who never used Marijuana
mean2=svyby( ~ irmjage , ~ irsex , subset(design.R, mrjflag==1),
svymean, na.rm = TRUE ); mean2
```

### Exhibit A.5.8 R Code (Calculates a Confidence Interval for Alcohol Drinker Prevalence and Estimated Totals Produced in Exhibit A.5.1)

This exhibit computes a 95 percent CI using the output data from [Exhibit A.5.1](#).

```
# define t-statistic
T_QNTILE=qt(c(.975), df=750); T_QNTILE

# Of various domains in Exhibit A.5.1, we only focus on estimates by
# year: YEAR1 and YEAR2.
##alcohol drinker proportion estimates by year
prop=svyby(~alcmon, ~year, design, svymean); prop
## weighted sample N by year
wdomain=svyby(~one, ~year, design, 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

# 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 A.5.9 R (Estimate Generation for Categorical Variable; i.e., Number of Days Used Substance in the Past Month among Past Month Users)

This exhibit shows how to compute estimates corresponding to levels of a categorical variable.

```
# Use the study design previously created from single year data in
# Exhibit A.5.7 FREQUENCY OF MARIJUANA USE BY PAST MONTH MARIJUANA
# USERS: 5 groups including non-users
# Sample total N of past month marijuana users
svyby( ~ one , ~ mrjmon, design=subset(design, mrjmon==1), unwtd.count
)
```

**Exhibit A.5.9 R (Estimate Generation for Categorical Variable; i.e., Number of Days Used Substance in the Past Month among Past Month Users) (continued)**

```
# weighted sample N in total
svytotal(~mrjmon, design=subset(design.R, mrjmon==1), na.rm = TRUE)

# Estimated N of past month marijuana users by 5 categories of
# marijuana use days
svytotal( ~ mrjmdays , design=subset(design.R, mrjmon==1), na.rm =
TRUE)

# Estimated % and SE by 5 categories of marijuana use days
b=svymean( ~ mrjmdays, design=subset(design.R, mrjmon==1), na.rm =
TRUE)
coef(b)*100 # percentage
SE(b)*100 # percentage SE
```

**Exhibit A.5.10 R Code (Statistical Tests of Differences between Two Groups when the Two Groups Overlap)**

This exhibit shows how to test among overlapping domains (i.e., scenarios where some cases are in both domains are being compared). A stacked dataset is created first that includes two records for each respondent in the overlap needed for analysis. The stacked data are then used to compute the test of differences among the overlapping domains. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```
# Significance testing of difference in cigarette user prevalence
# between full pop ages 18+ and full employed ages 18+
# Note that the two groups overlap and we only focus on one year

#Data management and create survey design
#1st data (indic=1) has employ=1 when IRWRKSTAT18 IN (1,2,3,4)
d.1=DATANAME_SINGLE
d.1$indic=1
d.1$employ=ifelse(d.1$irwrkstat18 %in% c(1,2,3,4), 1, 0)
#2nd data (indic=2) has employ=1 when IRWRKSTAT18 ==1
d.2= DATANAME_SINGLE
d.2$indic=2
d.2$employ=ifelse(d.2$irwrkstat18==1, 1, 0)
#append the two
d.1.2=rbind(d.1, d.2)

design.R <-
  svydesign(
    id = ~ verep ,
    strata = ~ vestr ,
    data = d.1.2 ,
    weights = ~ analwt2 ,
    nest = TRUE
  )

# Significance testing
svyttest(cigmon~indic, subset(design.R, catag18==1 & employ==1))
```

**Exhibit A.5.11 R Code (Tests of the Independence of the Prevalence Variable and Subgroup Variable)**

This exhibit 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. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```
# Chi-square test of independence: evaluate the association between
# two categorical variables

#prepare data
design.R <-
  svydesign(
    id = ~ verrep ,
      strata = ~ vestr ,
    data = DATANAME_SINGLE ,
    weights = ~ analwt2 ,
    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"))
)

# sample N of cigarette use
## cigarettes user / non-user
svyby( ~one, ~cigmon , subset(design.R, catag18==1), unwtd.count )
## N by employment status
svyby( ~one, ~irwrkstat18 , subset(design.R, catag18==1), unwtd.count
)
## cigarettes user by employment status
svyby(~one, ~irwrkstat18+cigmon, subset(design.R, catag18==1),
unwtd.count)

# weighted sample N
## cigarettes user by employment status
svytable(~irwrkstat18+cigmon, subset(design.R, catag18==1),
round=TRUE)
## % cigarettes user by employment status
svytable(~irwrkstat18+cigmon, subset(design.R, catag18==1),) %>%
prop.table(1)

# wald F test for independence between cigarettes use and employment
# status with 4 levels
a=svyloglin(~irwrkstat18+cigmon, subset(design.R, catag18==1))
b=update(a,~.^2); regTermTest(b, ~irwrkstat18:cigmon)
```

**Exhibit A.5.12 R Code (Pairwise Testing)**

Once an association is determined by a significant Wald  $F$  test computed in [Exhibit A.5.11](#), then pairwise testing as shown below can be done to determine individual significant tests across levels of employment status. DATANAME\_SINGLE is a dataset that has been subset to a single year of data.

```
# Create study design using previously created design from
# Exhibit A.5.11

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","part-
time"), 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("part-time",
"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"))
  )

# sample N involved in each of comparisons for Tukey test
## comparison: full-time vs part-time
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 )

# weighted sample N involved in each of comparisons for Tukey test
## comparison: full-time vs part-time
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 )

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

### Exhibit A.5.13 R Code (Test of Linear Trends Across Years)

This exhibit displays code that tests for a linear trend in past month alcohol use by sex across 4 years (YEAR) to evaluate change over time. Unlike the nonparametric approach used in SUDAAN ([Exhibit A.2.13](#)), this example uses the R function `svyglm` from the “survey” package, which is a model-based method. This method allows for testing changes over time while controlling for additional covariates. The calculated test statistic and  $p$  value for the model-based method will likely differ from the nonparametric method, but the overall interpretation of the results should be consistent. The input dataset has 4 years of data with a year variable defined for each year on the dataset.

```
# Read in 4 year data
FOURYEAR_DATA <- read_sas("YOUR DATA FILE")

design.trend <-
  svydesign(
    id = ~ verep ,
    strata = ~ vestr,
    data = FOURYEAR_DATA,
    weights = ~ analwt2,
    nest = TRUE)

design.trend <-
  update(design.trend,
         one = 1,
         year = factor(year, levels = 1:4, labels = c("2021", "2022",
"2023", "2024")),
         irsex = factor(irsex, levels = 1:2, labels = c("Male",
"Female")))

# Total Population
overall <- svyglm(alcmon~year, family=quasibinomial,
                 design=design.trend)
summary(overall)

# Males
male <- svyglm(alcmon~year, family=quasibinomial,
              subset(design.trend, irsex=="Male"))
summary(male)

# Females
female <- svyglm(alcmon~year, family=quasibinomial,
                subset(design.trend, irsex=="Female"))
summary(female)
```

## A.6 SPSS Exhibits

The SPSS exhibits in this section provide guidance on how to use various programming options to produce estimates for NSDUH using the statistical procedures documented within this report.

### Exhibit A.6.1 SPSS CSDESCRIPTIVES Procedure (Estimate Generation: Single Year and Pooled Years of Data)

This exhibit displays code that imports 2 years of data from SAS datasets into SPSS and creates a combined 2-year analysis dataset. In this process, it also computes a year indicator variable (YEAR) and a combined 2-year analytic weight (POOLED\_WEIGHT) that are used to compute the descriptive statistics means, sums, SEs, population size (POPSIZE), and design effect (DEFF) by survey year.

```
*Import SAS datafiles into SPSS for Year 1 Data.
```

```
GET  
SAS FILE=' [FILE PATH NAME] \DATANAME1.sas7bdat'.  
DATASET NAME DATA1 WINDOW=FRONT.
```

```
*Select key variables for analysis.
```

```
MATCH FILES /FILE=* /KEEP = QUESTID VESTR VEREP ANALWT2 IRSEX ALCMON  
MRJMDAYS MRJMON CIGMON IRWRKSTAT18 CATAG18.  
EXECUTE.
```

```
*Create YEAR variable for combining two years of data.
```

```
COMPUTE YEAR = 1.  
EXECUTE.
```

```
* Import SAS datafile into SPSS for Year 2 Data.
```

```
GET  
SAS DATA=' [FILE PATH NAME] \DATANAME2.sas7bdat'.  
DATASET NAME DATA2 WINDOW=FRONT.
```

```
*Select key variables for analysis.
```

```
MATCH FILES /FILE=* /KEEP = QUESTID VESTR VEREP ANALWT2 IRSEX ALCMON  
MRJMDAYS MRJMON CIGMON IRWRKSTAT18 CATAG18.  
EXECUTE.
```

```
*Create YEAR variable for combining two years of data.
```

```
COMPUTE YEAR = 2.  
EXECUTE.
```

```
*Create combined two-year dataset.
```

```
DATASET ACTIVATE DATA2.  
ADD FILES /FILE=*  
/FILE='DATA1'.  
EXECUTE.
```

```
DATASET NAME TWOYEAR_DATA.
```

### Exhibit A.6.1 SPSS CSDESCRIPTIVES Procedure (Estimate Generation: Single Year and Pooled Years of Data) (continued)

```

*Create two-year analysis weight for combined data.
COMPUTE POOLED_WEIGHT=ANALWT2 / 2.
EXECUTE.

DATASET CLOSE DATA1.

*Open dataset if closed (commented out in this example).
*GET
FILE='[FILE PATH NAME]\TWOYEAR_DATA.sav'.
*DATASET NAME TWOYEAR_DATA WINDOW=FRONT.

*Sort dataset by Nesting Variables.
SORT CASES BY VESTR(A) VEREP(A).

*Create Complex Sampling Plan necessary for estimating variance from a
complex sample.
* ID Nesting variables (VESTR and VEREP) and weight variable (ANALWT2
- 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.

*Single year analysis plan (commented out for this example).
*CSPLAN ANALYSIS
/PLAN FILE='[FILE PATH NAME]\SINGLE_PLAN.csplan'
/PLANVARS ANALYSISWEIGHT=ANALWT2
/SRSESTIMATOR TYPE=WR
/PRINT PLAN
/DESIGN STRATA=VESTR CLUSTER=VEREP
/ESTIMATOR TYPE=WR.

*Two year analysis plan.
CSPLAN ANALYSIS
  /PLAN FILE='[FILE PATH NAME]\TWOYEAR_PLAN.csplan'
  /PLANVARS ANALYSISWEIGHT=POOLED WEIGHT
  /SRSESTIMATOR TYPE=WR
  /PRINT PLAN
  /DESIGN STRATA=VESTR 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 A.6.1 SPSS CSDESCRIPTIVES Procedure (Estimate Generation: Single Year and Pooled Years of Data) (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 TWOYEAR_DATA.
CSDESCRIPTIVES
/PLAN FILE='FILE PATH NAME]\TWOYEAR_PLAN.csplan'
/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=' [FILE PATH NAME]\TWOYEAR_PLAN.csplan'
/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 A.6.2 SPSS Code Based on SPSS Output (Calculation of Standard Error of Totals for Fixed Domains)**

This exhibit computes the SE of the total for fixed domains using the data produced by the example in [Exhibit A.6.1](#). Because sex is a fixed domain, the SE of the totals would not be taken directly from the examples in [Exhibit A.6.1](#) but rather would be computed using the alternative SE estimation method as shown below.

*\*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 A.6.3 SPSS Code (Implementation of Suppression Rule)**

This exhibit applies the prevalence estimate suppression rule and the rule for means not bounded by 0 and 1 (commented out in the exhibit) using the data produced by the example in [Exhibit A.6.1](#). Starting with the 2024 NSDUH, the suppression rule was simplified, and the code for the new rules described in [Table 9.1](#) is shown in this exhibit.

*\*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.
```

```
DATASET ACTIVATE ALC_EST2.
```

```
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.
```

### Exhibit A.6.3 SPSS Code (Implementation of Suppression Rule) (continued)

```

*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)).
END IF.

*SUPPRESSION RULE FOR PREVALENCE ESTIMATES.
DO IF (SEMEAN=0).
    COMPUTE SUPRULE1=1.
END IF.
DO IF (RSELNP GT 0.175).
    COMPUTE SUPRULE2=1.
    END IF.
DO IF (NSUM <50).
    COMPUTE SUPRULE3=1.
    END IF.

DO IF (SUPRULE1 = 1 OR SUPRULE2 = 1 OR SUPRULE3 = 1).
    COMPUTE SUPPRESS = 1.
END IF.

*SUPPRESSION RULE FOR MEANS NOT BOUNDED BY 0 AND 1, I.E. AVERAGES
(COMMENTED OUT FOR THIS EXAMPLE).
*DO IF (RSE GT 0.5 OR NSUM < 10).
    *COMPUTE SUPAVGRULE=1.
*END IF.
DO IF (SUPAVGRULE=1).
    COMPUTE SUPPRESS=1.
END IF.
EXECUTE.
    
```



Released 2026

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 access and better outcomes for all.

1-877-SAMHSA-7 (1-877-726-4727) | 1-800-487-4889 (TDD) | [www.samhsa.gov](http://www.samhsa.gov)