2013 NATIONAL SURVEY ON DRUG USE AND HEALTH

METHODOLOGICAL RESOURCE BOOK SECTION 13: STATISTICAL INFERENCE REPORT

DISCLAIMER

SAMHSA provides links to other Internet sites as a service to its users and is not responsible for the availability or content of these external sites. SAMHSA, its employees, and contractors do not endorse, warrant, or guarantee the products, services, or information described or offered at these other Internet sites. Any reference to a commercial product, process, or service is not an endorsement or recommendation by SAMHSA, its employees, or contractors. For documents available from this server, the U.S. Government does not warrant or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed.

Substance Abuse and Mental Health Services Administration
Center for Behavioral Health Statistics and Quality
Rockville, Maryland

January 2015

This page intentionally left blank.

2013 NATIONAL SURVEY ON DRUG USE AND HEALTH: STATISTICAL INFERENCE REPORT

Prepared for the 2013 Methodological Resource Book (Section 13)

Contract No. HHSS283201000003C
RTI Project No. 0212800.001.208.001.002
Deliverable No. 41

RTI Authors:

Jeremy Aldworth
Stephanie Barnett
Devon Cribb
Teresa R. Davis
Misty S. Foster
Rachel Harter
Lisa E. Packer
Kathryn Spagnola

SAMHSA Authors:

Jonaki Bose
Art Hughes

RTI Project Director:

David Hunter

SAMHSA Project Officer:

Peter Tice

For questions about this report, please email Peter.Tice@samhsa.hhs.gov.

Prepared for Substance Abuse and Mental Health Services Administration, Rockville, Maryland

Prepared by RTI International, Research Triangle Park, North Carolina

January 2015

Recommended Citation: Center for Behavioral Health Statistics and Quality. (2015). *2013 National Survey on Drug Use and Health: Methodological Resource Book (Section 13, Statistical Inference Report)*. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Acknowledgments

This report was prepared for the Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality, by RTI International (a registered trademark and a trade name of Research Triangle Institute). Contributors to this report at RTI include Debbie Bond, James R. Chromy, Claudia Clark, Jennifer Cooney, Kristen Gulledge, and Philip Lee. SAMHSA reviewers include Sarra Hedden, Janet Kuramoto-Crawford, Pradip Muhuri, and Matthew Williams.

Table of Contents

Section

Appendix

This page intentionally left blank.

List of Tables

Table **Page**

List of Exhibits

Exhibit **Page**

List of Exhibits (continued)

Exhibit **Page**

List of Figures

Figure **Page**

1. Introduction

Statistical inference occurs whenever data obtained from sample observations belonging to and considered representative of a larger target population are used to make generalizations concerning the larger population. The target population for the 2013 National Survey on Drug Use and Health (NSDUH)[1] was the U.S. civilian, noninstitutionalized population aged 12 or older (at the time of their interview) in 2013. Measurements for this target population were the responses to the survey questions provided by people participating in the 2013 survey.

Statistical inferences concerning characteristics of interest for this population and various subpopulations are presented in the form of estimates derived from the sample data collected. Examples of the inferences made from the 2013 NSDUH data are presented in the 2013 detailed tables (Center for Behavioral Health Statistics and Quality [CBHSQ], 2014b) and the 2013 summary of national findings report (CBHSQ, 2014e) and include estimates of the number of people who were substance users during the past month, past year, and their lifetime, as well as the associated percentages (prevalence rates) of substance use for these reference periods. Inferences also were made for such categories as substance initiation; risk and protective factors; substance dependence, dependence or abuse, and treatment. Estimates of measures related to mental health problems are presented in the 2013 mental health detailed tables (CBHSQ, 2014c) and the 2013 mental health findings report (CBHSQ, 2014d).

The focus of this report is to describe the statistical inference procedures used to produce design-based estimates as presented in the 2013 detailed tables, the 2013 mental health detailed tables, the 2013 national findings report, and the 2013 mental health findings report.[2] The statistical procedures and information found in this report can also be generally applied to analyses based on the public use file as well as the restricted-use file available through the data portal.[3] This report is organized as follows: Section 2 provides background information concerning the 2013 NSDUH; Section 3 discusses the prevalence rates and how they were calculated, including specifics on topics such as mental illness, major depressive episode, and serious psychological distress; Section 4 briefly discusses how missing item responses of variables that are not imputed may lead to biased estimates; Section 5 discusses sampling errors and how they were calculated; Section 6 describes the degrees of freedom that were used when comparing estimates; and Section 7 discusses how the statistical significance of differences between estimates was determined. Section 8 discusses confidence interval estimation, and Section 9 describes how past year incidence of drug use was computed. Finally, Section 10 discusses the conditions under which estimates with low precision were suppressed. Appendix A contains examples that demonstrate how to conduct various statistical procedures documented

[1] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).

[2] Users of the 2013 public use file (CBHSQ/SAMHSA, 2014) may find inconsistencies in the variable names referenced in this report, Appendix A, the information presented in Table 5.1 in Section 5, and other specific numbers presented in this report (i.e., degrees of freedom). The specific information referenced in this report is based on the restricted-use dataset that was used to create the 2013 detailed tables (CBHSQ, 2014b), the 2013 mental health detailed tables (CBHSQ, 2014c), the 2013 mental health findings report (CBHSQ, 2014d), and the 2013 national findings report (CBHSQ, 2014e).

[3] The data portal can be found at http://www.icpsr.umich.edu/icpsrweb/content/SAMHDA/dataportal.html.

within this report using SAS® and SUDAAN® Software for Statistical Analysis of Correlated Data (RTI International, 2012) along with separate examples using Stata® software.

2. Background

The 2013 National Survey on Drug Use and Health (NSDUH) is an extension of a coordinated 5-year sample design providing estimates for all 50 States and the District of Columbia for the years 2005 through 2009, and then continuing through 2013. The survey is conducted using computer-assisted interviewing methods for the screening and interviewing of selected respondents. The respondent universe is the civilian, noninstitutionalized population aged 12 or older residing within the United States and the District of Columbia. People excluded from the universe include active-duty military personnel, people with no fixed household address (e.g., homeless and/or transient people not in shelters), and residents of institutional group quarters, such as correctional facilities, nursing homes, mental institutions, and long-term hospitals.

The coordinated design for 2005 through 2009 facilitated a 50 percent overlap in second-stage units (area segments) within each successive 2-year period from 2005 through 2009. The 2010–2013 NSDUHs continued the 50 percent overlap by retaining half of the second-stage units from the previous survey year. Those segments not retained from the previous year are considered "retired" from use; that is, these segments will not be used to field another main study sample.

Because the coordinated design enables estimates to be developed by State in all 50 States and the District of Columbia, States may be viewed as the first level of stratification as well as a reporting variable. Eight States were designated as large sample States (California, Florida, Illinois, Michigan, New York, Ohio, Pennsylvania, and Texas) with target sample sizes of 3,600. In 2013, sample sizes in these States ranged from 3,503 to 3,729 respondents. For the remaining 42 States and the District of Columbia, the target size was 900. Sample sizes in these States ranged from 852 to 953. State estimates combining multiple years of data and using either small area estimation[4] or direct estimation have been tabulated.

States were first stratified into a total of 900 State sampling regions (SSRs) (48 regions in each large sample State and 12 regions in each small sample State). These regions were contiguous geographic areas designed to yield on average the same number of interviews.[5] Unlike the 1999–2001 National Household Surveys on Drug Abuse and the 2002–2004 NSDUHs in which the first-stage sampling units were clusters of census blocks called area

[4] Small area estimation is a hierarchical Bayes modeling technique used to produce State-level estimates for a selected number of measures. For more details, see the *State Estimates of Substance Use and Mental Disorders from the 2009-2010 National Surveys on Drug Use and Health* (Hughes, Muhuri, Sathe, & Spagnola, 2012).

[5] Areas were defined using 2000 census geography. Dwelling units and population counts were obtained from the 2000 census data supplemented with revised population counts from Nielsen Claritas (see http://www.nielsen.com/us/en.html).

segments, the first stage of selection for the 2005–2013 surveys was census tracts.[6] This stage was included to contain sample segments within a single census tract to the extent possible.[7]

A total of 48 census tracts per SSR were selected, and within these sampled census tracts, adjacent census blocks were combined to form the second-stage sampling units or area segments. Although only 24 segments were needed to support the coordinated 5-year sample, an additional 24 segments were selected to support any supplemental studies that the Substance Abuse and Mental Health Services Administration may choose to field. These 24 segments constitute the reserve sample and were available for use in 2010, 2011, 2012, and 2013. Eight reserve sample segments per SSR were fielded during the 2013 survey year. Four of these segments were retained from the 2012 survey, and four were selected for use in the 2013 survey. These sampled segments were allocated equally into four separate samples, one for each 3-month period (calendar quarter) during the year. That is, a sample of addresses was selected from two segments in each calendar quarter so that the survey was essentially continuous in the field.

The overall design remained the same beginning with the 2002 NSDUH and continuing through the 2013 NSDUH. Survey respondents were given a $30 incentive payment for participation. Also, a pair-sampling strategy was implemented that increased the number of pairs selected in dwelling units with older people on the roster (Chromy & Penne, 2002). A dress rehearsal of the 2015 survey (to test revisions to respondent materials, questionnaire, equipment, and other changes) was conducted in late 2013 based on retired area segments used in the main survey earlier that year. More information about the dress rehearsal sample design can be found in the *2013 NSDUH Methodological Resource Book (MRB)* sample design report (Center for Behavioral Health Statistics and Quality [CBHSQ], 2014f).

During the 2008–2012 NSDUHs, a Mental Health Surveillance Study (MHSS) was embedded in the NSDUH. Each respondent in a subsample of adults (about 1,500 in 2008, 500 in 2009 and 2010, and 1,500 in 2011 and 2012) who had completed the NSDUH interview was administered the Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP) (First, Spitzer, Gibbon, & Williams, 2002).[8] The SCID was adapted for this study and was administered via paper and pencil over the telephone approximately 2 to 4 weeks after the NSDUH interview. In 2008, a split-sample MHSS was conducted to develop models using the SCID data that would use the Kessler-6 nonspecific psychological distress scale and two competing functional impairment scales to generate prevalence estimates of serious mental illness among adults aged 18 or older for the entire sample. Based on the results from the 2008 MHSS, a modified World Health Organization Disability Assessment Schedule (Rehm et al., 1999) was adopted for the 2009–2013 surveys. Using data from the 2008–2012 MHSS, an improved mental health prediction model was developed. The revised and improved model was used to predict mental illness starting with the 2012 NSDUH (more details on this model can be found in Section 3.1 of this report). For more

[6] Census tracts are relatively permanent statistical subdivisions of counties and provide a stable set of geographic units across decennial census periods.

[7] Some census tracts had to be aggregated in order to meet the minimum dwelling unit (DU) requirement of 150 DUs in urban areas and 100 DUs in rural areas.

[8] "DSM-IV-TR" stands for the *Diagnostic and Statistical Manual of Mental Disorders, 4th ed., Text Revision* (American Psychiatric Association, 2008).

information about the MHSS sample design, see the sample design report in the *2013 NSDUH MRB* (CBHSQ, 2014f).

The final respondent sample of 67,838 people for the 2013 NSDUH provides a sufficient sample to create domain estimates for a broad range of ages and other demographic categories. Individual observations are weighted so that the weighted sample represents the civilian, noninstitutionalized population aged 12 or older for the general U.S. population and for each of the individual States. The person-level weights in NSDUH are calibrated to population estimates (or control totals) obtained from the U.S. Census Bureau. For more information on the person-level sampling weight calibration in the *2013 NSDUH MRB*, see CBHSQ (2015b).

This page intentionally left blank.

3. Prevalence Rates

The national prevalence rates were computed using a multiprocedure package called SUDAAN® Software for Statistical Analysis of Correlated Data (RTI International, 2012). The final, nonresponse-adjusted, and poststratified analysis weights were used in SUDAAN to compute unbiased design-based drug use estimates. Appendix A contains examples that demonstrate how to compute the prevalence rates as defined below using SUDAAN (Exhibit A.1) as well as Stata® (Exhibit A.2).

Prevalence rates are the proportions of the population who exhibit characteristics of interest (such as substance use). Let $\hat{p}_d$ represent the prevalence rate of interest for domain *d*. Then $\hat{p}_d$ would be defined as the ratio

$$\hat{p}_d = \frac{\hat{Y}_d}{\hat{N}_d},$$

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, δ_i represents an indicator variable that is defined as 1 if the *i*th sample unit is in domain *d* and is equal to 0 otherwise, and y_i represents an indicator variable that 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 cases, estimates were produced from annual averages based on combined data from 2 or more survey years, and they are clearly labeled in the detailed tables. The data were combined for the 2010–2011, 2012–2013, and 2010–2013 surveys to obtain annual averages, and then the prevalence rates were computed in SUDAAN as described above. The annual averages were derived by concatenating the data for the respective years and dividing the analysis weights by a factor that varied depending on the number of years of concatenated data. The weight was divided by a factor of 2 for 2 years of concatenated data and a factor of 4 for 4 years of concatenated data.

3.1 Mental Illness

The Substance Abuse and Mental Health Services Administration (SAMHSA) has been publishing estimates of the prevalence of past year serious mental illness (SMI) and any mental illness (AMI) among adults aged 18 or older since the release of the 2008 National Survey on Drug Use and Health (NSDUH) national findings report (Office of Applied Studies, 2009b). Originally, estimates were based on a prediction model for mental illness developed using the 2008 data from the Mental Health Surveillance Study (MHSS) (referred to as the 2008 World Health Organization Disability Assessment Schedule [WHODAS] model). The 2008 NSDUH included a split sample, in which half the respondents (approximately 750) were administered the WHODAS and the other half the Sheehan Disability Scale (SDS). Two models were used to predict SMI for 2008, one for each impairment scale (WHODAS and SDS). The 2008 models for SMI were chosen so that estimates from the WHODAS and SDS samples were approximately equal; hence, SMI estimates for 2008 were based on both samples. The WHODAS model was determined to be a better predictor of SMI than the SDS model; therefore, starting in 2009, only the WHODAS impairment scale was administered in the NSDUH and used for estimating all levels of mental illness (SMI, mental illness [MI], AMI, mild mental illness [LMI], moderate mental illness [MMI], and serious or moderate mental illness [SMMI]).

Although SAMHSA continued to obtain clinical interviews after 2008, estimates of mental illness from the 2009, 2010, and 2011 NSDUHs have been based on the WHODAS model developed from the 2008 clinical assessment sample. The same model was applied to each year's NSDUH data to provide consistency in mental illness comparisons across the years. Producing a new model each year based on the small annual clinical samples (only 500 interviews in 2009 and 2010) would have resulted in large changes in the model parameters and corresponding prevalence rates due to sampling error, making it impossible to detect real trends in mental illness over time. Furthermore, an evaluation of the 2008 model, using the 2009 NSDUH clinical data, found that the model could not be significantly improved with the additional 500-case 2009 clinical sample. The clinical follow-up study, which started in 2008 and continued until 2012, led to a nationally representative sample of approximately 5,000 cases assigned to the WHODAS questions that were used to develop an improved mental illness prediction model (referred to as the 2012 WHODAS model). This revised and improved model was used to predict mental illness starting with the 2012 NSDUH and incorporates the NSDUH respondent's age and indicators of past year suicidal thoughts and depression, along with the variables that were specified in the 2008 model (e.g., variables for the Kessler-6 [K6] scale and the WHODAS), leading to more accurate estimates of mental illness (see below for details on the 2012 model and revised methodology).

For the 2012 and 2013 mental health detailed tables (Center for Behavioral Health Statistics and Quality [CBHSQ], 2013; 2014c), the 2008 and later year mental illness estimates were based on the revised model. As of October 2013, the 2008 detailed tables (Office of Applied Studies, 2009a) and the 2009–2011 mental health detailed tables (CBHSQ, 2010; 2012a; 2012c) containing estimates for past year mental illness for adults have been revised based on the 2012 model. The addition of these mental health predictors in the 2012 model, however, impacts the types of analyses that can be performed with the mental illness variables derived from the model. See the Using Mental Illness Variables in Analysis section below for more details. For detailed information on model revisions to the mental illness items, see Section B.4.3 in Appendix B of the *Results from the 2013 National Survey on Drug Use and Health: Mental Health Findings* (CBHSQ, 2014d). As with the mental illness estimates based on the 2008 model, the mental illness estimates based on the 2012 model are not comparable with SMI estimates produced from NSDUH data before 2004, and SMI estimates were not produced from 2004 to 2007; thus, long-term trend estimates are not available for SMI.

2012 SMI Prediction Model

The 2012 model is a prediction model for mental illness, and it was used to predict SMI and to estimate prevalence of SMI for the 2013 NSDUH. The prediction model is a weighted logistic regression. The response variable Y was defined so that $Y = 1$ when an SMI diagnosis was positive based on the clinical interview; otherwise, $Y = 0$. If $\mathbf{X}$ is a vector of realized explanatory variables, then the response probability $\pi = \Pr(Y = 1 \mid \mathbf{X})$ can be estimated using a weighted logistic regression model. For further technical details on the 2012 prediction models and the impact of the revised model on the 2008–2011 estimates, see the *2012 Mental Health Surveillance Study: Design and Estimation Report* (CBHSQ, 2014a) or Section B.4.3 in Appendix B of the *Results from the 2013 National Survey on Drug Use and Health: Mental Health Findings* (CBHSQ, 2014d).

The 2012 SMI prediction model was fit with data from 4,912 WHODAS MHSS respondents from 2008 through 2012, excluding one case from 2008 and one case from 2009 that were dropped because of data errors. The final WHODAS calibration model for the 2012 prediction model for SMI was determined as

$$\begin{aligned}\operatorname{logit}(\hat{\pi}) = \log[\hat{\pi} / (1-\hat{\pi})] = -5.972664 + 0.0873416X_k + 0.3385193X_w + 1.9552664X_s \\ + 1.1267330\,X_m + 0.1059137X_a \end{aligned} \qquad (1)$$

or

$$\hat{\pi} = \frac{1}{1+\exp[-(-5.972664 + 0.0873416X_k + 0.3385193X_w + 1.9552664X_s + 1.1267330\,X_m + 0.1059137X_a)},$$

where $\hat{\pi}$ refers to the estimate of the SMI response probability π. The covariates in equation (1) came from the main NSDUH interview data:

X_k = *Alternative Past Year K6 Score*: Past year K6 score of less than 8 recoded as 0; past year K6 score of 8 to 24 recoded as 1 to 17.

X_w = *Alternative WHODAS Score*: WHODAS item score of less than 2 recoded as 0; WHODAS item score of 2 to 3 recoded as 1, then summed for a score ranging from 0 to 8.

X_s = *Serious Thoughts of Suicide in the Past Year*: Coded as 1 if "yes"; coded as 0 otherwise.

X_m = *Past Year MDE*: Coded as 1 if the criteria for past year MDE were met;[9] coded as 0 otherwise.

X_a = *Recoded Age*: Coded as age minus 18 if aged 18 to 30; coded as 12 otherwise.

[9] In this situation, the past year MDE measure is from the main NSDUH interview (i.e., not from the Structured Clinical Interview for DSM-IV). See section B.4.4 of the 2013 mental health findings report (CBHSQ, 2014d).

A cut point probability π_0 was determined, so that if $\hat{\pi} \geq \pi_0$ for a particular respondent, then he or she was predicted to be SMI positive; otherwise, he or she was predicted to be SMI negative. The cut points were chosen so that the weighted numbers of false positives and false negatives in the MHSS dataset were as close to equal as possible. The predicted SMI status for all adult NSDUH respondents was used to compute prevalence estimates of SMI. In the 2012 SMI WHODAS prediction model, the respondent is classified as having past year SMI if the predicted probability of SMI is greater than or equal to 0.260573529 (SMI cutoff point). See Table 3.1 for the model specifications. Table 3.2 contains the cutoff points for other mental illness levels.

Modified 2012 Model for the 2008 SDS Half Sample

As noted previously, the 2008 NSDUH data included a split sample. Similar to the 2008 model, the revised 2012 model also has an alternative model for the SDS data that was fit with data from the complete 2008–2012 MHSS clinical sample that contains 5,653 MHSS respondents, excluding 4 cases from 2008 (one from the WHODAS half sample and three from the SDS half sample) and 1 case from 2009 that were dropped because of data errors.

The modified 2012 SMI prediction model for the SDS half sample was

$$\begin{aligned}\text{logit}(\hat{\pi}) = \log[\hat{\pi}/(1-\hat{\pi})] = -5.7736246 + 0.1772067X_k + 1.8392433X_s \\ + 1.6428623X_m + 0.1231266X_a \end{aligned} \qquad (2)$$

or

$$\hat{\pi} = \frac{1}{1+\exp[-(-5.7736246 + 0.1772067X_k + 1.8392433X_s + 1.6428623X_m + 0.1231266X_a)}.$$

All of the covariates in equation (2) also appeared in equation (1).

Similar to the WHODAS model, a cut point probability π_0 was determined, so that if $\hat{\pi} \geq \pi_0$ for a particular respondent, then he or she was predicted to be SMI positive; otherwise, he or she was predicted to be SMI negative. The cut points were chosen so that the weighted numbers of false positives and false negatives in the MHSS dataset were as close to equal as possible. In the 2012 SMI SDS half sample prediction model, the respondent is classified as having past year SMI if the predicated probability of SMI is greater than or equal to 0.236434 (SMI cutoff point). Although the SDS half sample prediction model was fit across all years and the cutoff points were determined based on all years, the cutoff points were used only for the main study respondents in the 2008 sample B to predict the SMI positives. See Tables 3.1 and 3.2.

Table 3.1 Final SMI Prediction Models in the 2008–2012 MHSS

	Beta	Beta SE	*T* Statistic	*P* Value	*DF*	Wald *P* Value[1]
WHODAS Sample (2008A-2012)						
Intercept	-5.9726640	0.3201	-18.6586	0.0000		
Alt PY K6	0.0873416	0.0248	3.5247	0.0009	1	0.0009
Alt WHODAS	0.3385193	0.0349	9.7034	0.0000	1	0.0000
PY Suicidal Thoughts	1.9552664	0.2164	9.0342	0.0000	1	0.0000
PY MDE	1.1267330	0.2196	5.1308	0.0000	1	0.0000
Age1830	0.1059137	0.0244	4.3380	0.0001	1	0.0001
WHODAS and SDS Samples (2008–2012)[2]						
Intercept	-5.7736246	0.3479	-16.5960	0.0000		
Alt PY K6	0.1772067	0.0190	9.3251	0.0000	1	0.0000
PY Suicidal Thoughts	1.8392433	0.1941	9.4781	0.0000	1	0.0000
PY MDE	1.6428623	0.2119	7.7528	0.0000	1	0.0000
Age1830	0.1231266	0.0259	4.7482	0.0000	1	0.0000

Age1830 = recoded age variable; Alt = alternative; *DF* = degrees of freedom; K6 = Kessler-6, a six-item psychological distress scale; MDE = major depressive episode; MHSS = Mental Health Surveillance Study; PY = past year; SDS = Sheehan Disability Scale; SE = standard error; SMI = serious mental illness; WHODAS = eight-item World Health Organization Disability Assessment Schedule; 2008A = 2008 WHODAS half sample.

[1] The *p* value is obtained from the overall model fitting.

[2] The model is fit over the WHODAS and SDS samples in 2008–2012 but is used only to produce predictions for the 2008 SDS sample.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2008–2012.

Table 3.2 Cut Point Probabilities for SMI, AMI, and SMMI, by 2012 Model

	Cut Point Probability
WHODAS Sample (2008A-2012)	
SMI	0.260573529
AMI	0.0192519810
SMMI	0.077686285365
WHODAS and SDS Samples (2008–2012)[1]	
SMI	0.236434
AMI	0.019182625
SMMI	0.06616398

AMI = any mental illness; SDS = Sheehan Disability Scale; SMI = serious mental illness; SMMI = serious or moderate mental illness; WHODAS = World Health Organization Disability Assessment Schedule; 2008A = 2008 WHODAS half sample.

[1] The model is fit over the WHODAS and SDS samples in 2008–2012, but the cut point predictions are only used to produce predictions for the 2008 SDS sample.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2008–2012.

Weights

For the 2008 NSDUH, although SMI data for both half samples (SDS and WHODAS) could be analyzed together when using the 2008 model, the AMI, SMMI, LMI, and MMI data from the two half samples could not be combined for analysis. Under the 2012 model, both the 2008 half samples can be combined to analyze SMI and the other levels of mental illness because the 2012 models were generated so the estimates would be comparable.

Mental illness measures (i.e., SMI, AMI, SMMI, MMI, and LMI) that are defined based on the 2012 model should be analyzed using the standard analysis weight, ANALWT, for all survey years 2008 through 2013. With the revised 2012 model, both the WHODAS and SDS 2008 half samples can be combined to form single estimates and use ANALWT.

This differs from the initial recommendation for analyzing measures of mental illness besides SMI based on the 2008 model. Because of the 2008 split sample, an adjusted mental health sample weight, MHSAMPWT, was created so that the WHODAS and SDS half samples were separately representative of the civilian, noninstitutionalized population aged 18 or older. However, this weight should not be used to analyze 2008 mental illness data based on the 2012 model.

Standard Errors for Mental Illness Estimates

For the 2013 mental health detailed tables and the mental health findings report (CBHSQ, 2014c; 2014d), standard errors (SEs) for mental illness estimates (SMI, AMI, SMMI, MMI, and LMI) were computed using the NSDUH dichotomous variable values without taking into account any variance introduced through using a model based on the clinical subsample data. This ignores the added error resulting from fitting the 2012 SMI model, which can be very large. See the *2012 Mental Health Surveillance Study: Design and Estimation Report* (CBHSQ, 2014a) for details. These *conditional* SEs (conditional on the model predictions being correct) are useful when making comparisons across years and across subpopulations within years because the errors due to model fitting are nearly the same across the estimates being compared, and consequently, they roughly cancel each other out.

Using Mental Illness Variables in Analysis

The mental illness measures (i.e., SMI, AMI, SMMI, MMI, and LMI) that were defined based on the 2012 model were examined to determine how they were associated with the mental health predictor variables in the 2012 model. It was found that the 2012 model significantly overestimated the proportion of adults aged 18 or older with SMI (and those with AMI) who had suicidal thoughts in the past year and also the proportion of adults who had MDE in the past year (as compared with the clinical interview estimates of the same categories). Therefore, it is recommended that the mental illness measures derived from the 2012 model should not be used when analyzing past year suicidal thoughts, past year MDE, or other associated variables (including past year suicide attempts, suicide plans, medical treatment for suicide attempts, or lifetime MDE). Similarly, it is recommended that model-based mental illness measures should not be used in conjunction with the K6 variables (including serious psychological distress) or WHODAS variables in any analyses (CBHSQ, 2014a).

3.2 Adult Major Depressive Episode (MDE)

The past year adult MDE estimates shown in the 2013 mental health detailed tables (CBHSQ, 2014c) are based on the full sample as was done in the 2010–2012 mental health detailed tables (CBHSQ, 2012a; 2012c; 2013). This differs from the 2008 past year MDE estimates shown in both the 2008 detailed tables (Office of Applied Studies, 2009a) and the 2009 mental health detailed tables (CBHSQ, 2010), which were based only on the sample of adult respondents who received the WHODAS questions in the mental health questionnaire module that preceded the adult depression questionnaire module. The analysis of 2008 MDE data was restricted to only the WHODAS half sample because of apparent reporting differences (context effects) between the half sample that was administered the WHODAS and the other half sample of adult respondents who received the SDS questions (Dean & LeBaron, 2009). Both half samples have issues with context effects not seen in 2007 and previous years due to the revisions to the mental health module preceding the adult depression module. To address the break in comparability of the adult MDE data beginning in 2008 and to estimate adult MDE based on the full sample of adults from 2008, adjusted versions of lifetime and past year MDE variables for adults were created retroactively for 2005 to 2008. These variables were adjusted to make MDE estimates from the SDS half sample in 2008 and from all adult respondents for 2005 to 2007 comparable with the MDE estimates based on data from the half sample that received the WHODAS in 2008 and from all adult respondents in later years (2009 onward). The adjusted data from 2005 to 2008 can be used in conjunction with unadjusted data from later years to estimate trends in adult MDE over the entire period from 2005 to 2013.

In the 2013 mental health detailed tables (CBHSQ, 2014c), ANALWT was used to generate all estimates of adult MDE. More information about how the statistically adjusted adult MDE variables were created can be found in Section B.4.4 in Appendix B of the 2013 mental health findings report (CBHSQ, 2014d) and in the report describing the adjustments (Aldworth, Kott, Yu, Mosquin, & Barnett-Walker, 2012).

3.3 Serious Psychological Distress (SPD)

The K6 scale, a measure of psychological distress, was used to create the SPD variable. Before 2008, the K6 consisted of one set of questions that asked adult respondents about symptoms of psychological distress in the month when they were the most depressed, anxious, or emotionally distressed in the past year. Starting in 2008, the K6 consisted of two sets of questions that asked adult respondents how frequently they experienced symptoms of psychological distress during two different periods: (1) during the past 30 days, and (2) if applicable, the month in the past year when they were at their worst emotionally. Respondents were asked about this second period only if they indicated that there was a month in the past 12 months when they felt more depressed, anxious, or emotionally stressed than they felt during the past 30 days. Because of this change, past year K6 and SPD estimates from years before 2008 were no longer comparable with estimates from 2008 onward. To address this comparability issue, adjusted versions of the past year worst K6 total score and past year SPD variables were created for each of the years from 2005 to 2007 to make the 2005–2007 past year K6 scores and past year SPD estimates comparable with their 2008–2013 counterparts.

In the 2013 mental health detailed tables (CBHSQ, 2014c), ANALWT was used to generate 2005–2013 estimates of past year SPD and 2008–2013 estimates of past month SPD. The 2013 mental health findings report (CBHSQ, 2014d) did not present SPD estimates. More information about how the adjusted K6 and SPD variables were created can be found in the report describing these adjustments (Aldworth et al., 2012).

3.4 Decennial Census Effects on NSDUH Substance Use and Mental Health Estimates

As discussed in Section 2, the person-level weights in NSDUH were calibrated to population estimates (or control totals) obtained from the U.S. Census Bureau. For the weights in 2002 through 2010, annually updated control totals based on the 2000 census were used.[10] Beginning with the 2011 weights, however, the control totals from the U.S. Census Bureau are based on the 2010 census. Two investigations were implemented at the national level to assess the effects of using control totals based on the 2010 census instead of the 2000 census. One of these investigations focused specifically on measures of substance use that are used in the 2011 national findings report (CBHSQ, 2012e) and detailed tables (CBHSQ, 2012b), while a separate analysis was conducted to evaluate the impact of the weighting changes on mental health estimates in the 2011 mental health findings report (CBHSQ, 2012d) and associated mental health detailed tables (CBHSQ, 2012c). Because both the 2012 and 2013 NSDUH estimates are based on weights that were poststratified to population control totals that were in turn based on projections from the 2010 census, 2-year trend comparisons between 2012 and 2013 are not subject to census effects. However, trends between 2010 (or earlier years) and 2011 (or later years) may be influenced by census effects, especially for particular subgroups (e.g., people reporting two or more races for both investigations, people reporting American Indian or Alaska Native or Native Hawaiian or Other Pacific Islander). An additional investigation was done at the State level to evaluate the impact of census effects on model-based small area estimates (SAEs).

For more information on the impact of decennial census effects on NSDUH substance use direct estimates, see Section B.4.3 in Appendix B of the 2011 national findings report (CBHSQ, 2012e). For more information on the impact of decennial census effects on NSDUH mental health direct estimates, see Appendix A of the 2011 mental health findings report (CBHSQ, 2012d). For more information on the impact of the decennial census effects on NSDUH model-based SAEs, see http://www.samhsa.gov/data/NSDUH/2k12State/NSDUHsae2012/Index.aspx. Additionally, for more information on the sampling weight calibration in the 2011 NSDUH, see the person-level sampling weight calibration report (Chen et al., 2013).

3.5 Using Revised Estimates for 2006 to 2010

During regular data collection and processing checks for the 2011 NSDUH, data errors were identified. These errors affected the data for Pennsylvania (2006–2010) and Maryland

[10] In addition to the standard 2010 analysis weights poststratified to 2000 census control totals, special weights that were poststratified to 2010 census control totals are available on the 2010 NSDUH public use file (CBHSQ/SAMHSA, 2012).

(2008–2009). Cases with erroneous data were removed from the data files, and the remaining cases were reweighted to provide representative estimates. The errors had minimal impact on the national estimates and no effect on direct estimates for the other 48 States and the District of Columbia. In reports where model-based small area estimation techniques were used, estimates for all States may have been affected, even though the errors were concentrated in only two States. In reports that did not use model-based estimates, the only estimates appreciably affected are estimates for Pennsylvania, Maryland, the mid-Atlantic division, and the Northeast region. The 2013 detailed tables (CBHSQ, 2014b), the 2013 mental health detailed tables (CBHSQ, 2014c), the 2013 mental health findings report (CBHSQ, 2014d), and the 2013 national findings report (CBHSQ, 2014e) did not include State-level or model-based estimates. However, they did include estimates for the mid-Atlantic division and the Northeast region. Estimates based on 2006–2010 data may differ from previously published estimates. Tables and estimates based only on 2011 or later data are unaffected by these data errors. All affected tables, that is, tables with estimates based on 2006–2010 data, contain a note to indicate this to the user.

Caution is advised when comparing data from older reports with data from more recent reports that are based on corrected data files. As discussed above, comparisons of estimates for Pennsylvania, Maryland, the mid-Atlantic division, and the Northeast region are of most concern, whereas comparisons of national data or data for other States and regions are essentially still valid. A selected set of corrected versions of reports and tables have been produced. In particular, a set of modified detailed tables that include revised 2006–2010 estimates for the mid-Atlantic division and the Northeast region for certain key measures have been released. Given the change noted above, comparisons between unrevised 2006–2010 estimates and estimates based on 2011–2013 data for the areas of most concern are not recommended.

This page intentionally left blank.

4. Missingness

4.1 Potential Estimation Bias Due to Missingness

In the 2013 National Survey on Drug Use and Health (NSDUH), many variables, including core drug and demographic variables, had missing item response values imputed. See the 2013 NSDUH editing and imputation report (Center for Behavioral Health Statistics and Quality [CBHSQ], 2015a) for further details. However, the missing item responses of many other variables were not imputed, and these missing responses may lead to biased estimates in the 2013 detailed tables (CBHSQ, 2014b) and the 2013 mental health detailed tables (CBHSQ, 2014c). In addition, another source of potential uncertainty about some estimates may occur because of the way unknown item responses (e.g., blank, "don't know," "refused") were actually coded for different variables. For example, some recoded variables (i.e., variables created from one or more source variables) classified unknown item responses in the source variable(s) as missing values, whereas others did not. See Ruppenkamp, Emrich, Aldworth, Hirsch, and Foster (2006) for further details.

Recall from Section 3 that prevalence rates are defined as the proportions of the population who exhibit characteristics of interest. Let $\hat{p}_d$ represent the estimated prevalence rate of interest for domain *d*, with $\hat{p}_d$ defined as

$$\hat{p}_d = \frac{\hat{Y}_d}{\hat{N}_d},$$

where $\hat{Y}_d$ = estimated number of people exhibiting the characteristic of interest in domain *d*, and $\hat{N}_d$ = estimated population total for domain *d*.

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., receipt of past year mental health treatment/counseling) 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 cases, the estimates $\hat{N}_d$ 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

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

where $\hat{N}$ = estimated population total, $\hat{N}_d$ = estimated population total for domain *d*, $d = 1, 2, ..., D$, 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 of $\hat{N}_m$ and $\hat{Y}_m$ is positive, then $\hat{p}_d$ may be biased by some unknown amount.

In the 2013 detailed tables (CBHSQ, 2014b) and the 2013 mental health detailed tables (CBHSQ, 2014c), 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 were based on or excluded missing values. This problem may be illustrated by the following example, which corresponds to information presented in Tables 2.9A and 2.9B of the 2013 mental health detailed tables.

Mental health Table 2.9A presents estimates of the past year use of several types of illicit drugs among people aged 12 to 17 for 2012 and 2013. These analysis variables are grouped into a two-level domain variable that is categorized according to whether a respondent had a past year major depressive episode (MDE). In 2013, mental health Table 3.2A shows the population estimate of people aged 12 to 17 as approximately 24,893,000. However, the subdomain population estimates summed to approximately 24,285,000, resulting in an estimate of $\hat{N}_m$ = 607,000 (approximately 2.4 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." In 2013, the total estimate of people aged 12 to 17 who used illicit drugs in the past year was approximately 4,287,000. However, the estimates of people aged 12 to 17 who used illicit drugs in the past year among the valid subdomains (where past year MDE status was not missing) summed to 4,134,000, resulting in an estimate of $\hat{Y}_m$ = 153,000 (approximately 3.6 percent of the total population).

Mental health Table 2.9B presents prevalence estimates of the past year use of several types of illicit drugs among people aged 12 to 17 for 2012 and 2013. Because $\hat{N}_m$ is positive and $\hat{Y}_m$ is positive for the analysis variable, "Illicit Drugs," the prevalence estimates for this variable may be biased by some unknown amount across the two domains. The 2013 prevalence estimates reported in mental health Table 2.9B for youths who had or did not have past year MDE are 33.2 and 15.1 percent, respectively. It can be shown that the approximate range of possible bias values for each of these estimates is as follows: between -4.96 and 3.73 percent and between -0.31 and 0.59 percent, respectively.

4.2 Variance Estimation in the Presence of Missingness

SUDAAN® Software for Statistical Analysis of Correlated Data (RTI International, 2012) uses the number of strata 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 there may be individuals in the target population who have nonmissing values in PSUs where no sample members have nonmissing values.

To illustrate how this is operationalized in SUDAAN, consider the following example. Suppose there is interest in calculating the mean of some variable (say, X), but there are missing values associated with X. SUDAAN then creates an internal subpopulation indicator variable (say, δ), where $\delta = 1$ if X is not missing, and $\delta = 0$ if X is missing. SUDAAN then internally calculates the mean and variance of X by using δX, assuming 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, and this 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.

This page intentionally left blank.

5. Sampling Error

As were the prevalence rates, all of the variance estimates for prevalence (including those for prevalence based on annual averages from combined data) were calculated using a method in SUDAAN® that is unbiased for linear statistics. This method is based on multistage clustered sample designs where the first-stage (primary) sampling units are drawn with replacement.

Because of the complex nature of the sampling design for the National Survey on Drug Use and Health (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. Starting with the 2005 NSDUH,[11] a change was made in the way the key nesting variables were defined. Each State sampling region (SSR) appears in a different variance estimation stratum every quarter. This method had 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.

Two replicates per year are defined within each variance stratum (VEREP). Each variance replicate 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, thus constituting the 50 percent overlap between survey years. A segment stays in the same VEREP for the 2 years it is in the sample. This simplifies computing standard errors (SEs) for estimates based on combined data from adjacent survey years.

Although the SEs of estimates of means and proportions can be calculated appropriately in SUDAAN using a Taylor series linearization approach, SEs of estimates of totals may be underestimated in situations where the domain size is poststratified to data from the U.S. Census Bureau. Because of this underestimation, alternatives for estimating SEs of totals were implemented in all of the 2013 detailed tables (Center for Behavioral Health Statistics and Quality [CBHSQ], 2014b) and the 2013 mental health detailed tables (CBHSQ, 2014c), where appropriate.

Estimates of means or proportions, $\hat{p}_d$, such as drug use prevalence rates for a domain *d*, can be expressed as a ratio estimate:

$$\hat{p}_d = \frac{\hat{Y}_d}{\hat{N}_d},$$

where $\hat{Y}_d$ is a linear statistic estimating the number of substance users in the domain *d*, and $\hat{N}_d$ is a linear statistic estimating the total number of people in domain *d* (both users and nonusers).

[11] The new design variables were created retroactively for 1999 through 2004; however, the old design variables continue to be used to generate 2002–2004 estimates in multiyear trend detailed tables and mental health detailed tables for consistency with previously published estimates. Analyses beyond the detailed tables and mental health detailed tables typically use the new design variables for all available years.

The SUDAAN software package is used to calculate direct estimates of $\hat{Y}_d$ and $\hat{N}_d$ and also 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 free of sampling error, an appropriate estimate of the SE for the total number of substance users is

$$\text{SE}(\hat{Y}_d) = \hat{N}_d \text{SE}(\hat{p}_d).$$

This approach is theoretically correct when the domain size estimates, $\hat{N}_d$, are among those forced to match their respective U.S. Census Bureau population estimates through the weight calibration process. In these cases, $\hat{N}_d$ is not subject to a sampling error induced by the NSDUH design. For more information on the person-level sampling weight calibration in the 2013 NSDUH, see CBHSQ (2015b).

For estimated domain totals, $\hat{Y}_d$, where $\hat{N}_d$ is not fixed (i.e., where domain size estimates are not forced to match the U.S. Census Bureau population estimates), this formulation 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 most estimates in this study.

For various subsets of estimates, the above approach yielded an underestimate of the variance of a total because $\hat{N}_d$ was subject to considerable variation. In 2000, an approach was implemented to reflect more accurately the effects of the weighting process on the variance of total estimates. This approach consisted of calculating SEs of totals for all estimates in a particular detailed table using the formula above when a majority of estimates in a table were among domains in which $\hat{N}_d$ was fixed during weighting or if it could be assumed that the sampling variation in $\hat{N}_d$ was negligible. Detailed tables in which the majority of estimates were among domains where $\hat{N}_d$ was subject to considerable variability were calculated directly in SUDAAN.

To improve on the accuracy of the SEs, a "mixed" method approach was implemented in which tables might include more than one method of SE estimation. This mixed approach was applied to selected tables in the 2004 NSDUH, and it was implemented across all tables starting with the 2005 NSDUH and continuing in the 2013 NSDUH. 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 formula above, and all other estimates were calculated directly in SUDAAN, regardless of other estimates within the same table. The set of domains considered controlled (i.e., those with a fixed $\hat{N}_d$) was restricted to main effects and two-way interactions to maintain continuity between years. Domains consisting of three-way interactions may be controlled in 1 year but not necessarily in preceding or subsequent years. The use of such

SEs 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 SAS®, SUDAAN, and Stata® code examples that demonstrate how to compute SEs of proportions as well as both types of SEs of totals (controlled or uncontrolled; see Exhibits A.1 to A.4).

Table 5.1 contains a list of domains with a fixed $\hat{N}_d$ for the restricted use data file.[12] This table includes both the main effects and two-way interactions and may be used to identify the method of SE calculation employed for estimates of totals in the 2013 detailed tables (CBHSQ, 2014b) and the 2013 mental health detailed tables (CBHSQ, 2014c). For example, Table 1.23 of the 2013 detailed tables presents estimates of illicit drug use among people aged 18 or older within the domains of gender, Hispanic or Latino (referred to as "Hispanic" hereafter) origin and race, education, and current employment. Estimates among the total population (age main effect), males and females (age by gender interaction), and Hispanics and non-Hispanics (age by Hispanic origin interaction) were treated as controlled in this table, and the formula above was used to calculate the SEs. The SEs for all other estimates, including white and black or African American (age by Hispanic origin by race interaction), were calculated directly from SUDAAN. It is important to note that estimates presented in the 2013 detailed tables and 2013 mental health detailed tables for racial groups are among non-Hispanics, unless noted otherwise. For instance, the domain for whites is actually non-Hispanic whites and is therefore a two-way interaction. Although not reported on in the 2013 detailed tables or the 2013 mental health detailed tables, additional geographic interactions are also treated as domains with fixed $\hat{N}_d$ for other NSDUH analyses. Similar to geographic region and division, a State is considered a controlled domain, and two-way interactions with State and gender, Hispanic origin, quarter, and age group (12-17, 18-25, and 26 or older) are all treated as domains with fixed $\hat{N}_d$.

[12] See the estimation of totals section in the 2013 public use data file introduction for a list of domains with fixed $\hat{N}_d$ (CBHSQ/SAMHSA, 2014).

Table 5.1 Demographic and Geographic Domains Forced to Match Their Respective U.S. Census Bureau Population Estimates through the Weight Calibration Process, 2013

Main Effects	Two-Way Interactions
Age Group	
12-17	
18-25	**Age Group** × **Gender**
26-34	(e.g., Males Aged 12 to 17)
35-49	
50-64	
65 or Older	**Age Group** × **Hispanic Origin**
All Combinations of Groups Listed Above[1]	(e.g., Hispanics or Latinos Aged 18 to 25)
Gender	
Male	
Female	**Age Group** × **Race**
Hispanic Origin	(e.g., Whites Aged 26 or Older)
Hispanic or Latino	
Not Hispanic or Latino	
Race	**Age Group** × **Geographic Region**
White	(e.g., People Aged 12 to 25 in the Northeast)
Black or African American	
Geographic Region	
Northeast	**Age Group** × **Geographic Division**
Midwest	(e.g., People Aged 65 or Older in New England)
South	
West	
Geographic Division	**Gender** × **Hispanic Origin**
New England	(e.g., Not Hispanic or Latino Males)
Middle Atlantic	
East North Central	
West North Central	**Hispanic Origin** × **Race**
South Atlantic	(e.g., Not Hispanic or Latino Whites)
East South Central	
West South Central	
Mountain	
Pacific	

[1] Combinations of the age groups (including but not limited to 12 or older, 18 or older, 26 or older, 35 or older, and 50 or older) also were forced to match their respective U.S. Census Bureau population estimates through the weight calibration process.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2013.

6. Degrees of Freedom

To determine whether the observed difference between estimates is statistically significant, the degrees of freedom (*df*) are needed to locate the corresponding probability level (*p* value) of the test statistic. 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 degrees of freedom 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 degrees of freedom 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 degrees of freedom of a parameter estimate is 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 degrees of freedom are also used to compute the confidence intervals (CIs) discussed in Section 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 degrees of freedom.

Starting in 2005, there was a change in definition to the variance estimation strata for the National Survey on Drug Use and Health (NSDUH; see footnote 11). This change in definition, which was applied to the 2005–2013 NSDUHs, had the effect of increasing the number of degrees of freedom for State-level estimates while preserving the number of degrees of freedom for national estimates (900). The degrees of freedom are calculated as the number of primary sampling units (variance replicates) minus the number of strata for the data being analyzed. Because the NSDUH sample design provides for estimates by State in all 50 States plus the District of Columbia, States may be viewed as the first level of stratification. When producing NSDUH estimates on the national level, including estimates based on annual averages from combined data, there are 900 degrees of freedom. If an analysis involves only certain States, the degrees of freedom change depending on whether the State is a large sample or small sample State. The large sample States (i.e., California, Florida, Illinois, Michigan, New York, Ohio, Pennsylvania, and Texas) have 192 degrees of freedom because each large State is in 192 strata. All of the other States (i.e., the small sample States, which include the District of Columbia) have 48 degrees of freedom because each small State is in 48 different strata. Note that the 2013 detailed tables (CBHSQ, 2014b) and the 2013 mental health detailed tables (CBHSQ, 2014c) use 900 degrees of freedom for all estimates, including those for geographic regions and divisions.

When testing between groups that have differing degrees of freedom, the minimum of the degrees of freedom from the groups is used, which makes for a more conservative test. For all nonstandard geographic areas (e.g., geographic areas other than region and division) or when multiple years are grouped for analyses, the degrees of freedom are calculated outside of SUDAAN®, and this value is entered manually into SUDAAN for use in testing. Appendix A contains an example that demonstrates how to define the degrees of freedom within SUDAAN (RTI International, 2012) or Stata® to compute design-based estimates.

For an analysis of a group of States, the degrees of freedom would be less than or equal to the sum of the degrees of freedom for each individual State due to overlap of strata. The specific number of degrees of freedom can be computed by counting the unique values of VESTR (variance estimation [pseudo] stratum) for the particular geographic area of interest. For these types of specific State analyses (or other subpopulations of interest), the degrees of freedom can be specifically indicated in SUDAAN (RTI International, 2012); otherwise, the degrees of freedom are computed using the entire dataset. Similar methods can be used to compute appropriate degrees of freedom for any geographic region comprising counties or States as well. As noted previously, the detailed tables and mental health detailed tables deviate from this rule and use 900 degrees of freedom for all estimates (including census region and division estimates and estimates for all years including pooled years). The technique of counting the number of unique values of VESTR can also be used for analyses combining survey data across years.

7. Statistical Significance of Differences

Once the degrees of freedom have been determined, various methods used to compare prevalence estimates may be employed. This section describes some of these methods. Customarily, the observed difference between estimates is evaluated in terms of its statistical significance. Statistical significance is based on the *p* value of the test statistic and refers to the probability that a difference as large as that observed would occur due to random variability in the estimates if there were no differences in the prevalence rates being compared. The significance of observed differences is generally reported at the .05 and .01 levels when the *p* value is defined as less than or equal to the designated significance level.

Significance tests were conducted on differences between prevalence estimates from the 2013 National Survey on Drug Use and Health (NSDUH) and previous years of NSDUH back to 2002. Due to survey design changes implemented in 2002, data from the 2002–2013 NSDUHs should not be compared with data from earlier survey years. Significance tests also were conducted on differences of prevalence estimates between combined 2010–2011 survey data and combined 2012–2013 survey data. Within-year tests were conducted on differences between prevalence estimates for various populations (or subgroups) of interest using data from the 2013 survey.

When comparing prevalence estimates, one can test the null hypothesis (no difference between rates) against the alternative hypothesis (there is a difference in prevalence rates) using the standard *t* test (with the appropriate degrees of freedom) for the difference in proportions test, expressed as

$$\mathrm{t}_{df} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\mathrm{var}(\hat{p}_1) + \mathrm{var}(\hat{p}_2) - 2\mathrm{cov}(\hat{p}_1, \hat{p}_2)}}, \tag{1}$$

or

$$\mathrm{t}_{df} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\mathrm{var}(\hat{p}_1) + \mathrm{var}(\hat{p}_2) - 2\,\rho(\hat{p}_1, \hat{p}_2)\mathrm{SE}(\hat{p}_1)\mathrm{SE}(\hat{p}_2)}} \tag{2}$$

where in both formulas *df* = the appropriate degrees of freedom, $\hat{p}_1$ = the first prevalence estimate, $\hat{p}_2$ = the second prevalence estimate, $\mathrm{var}(\hat{p}_1)$ = the variance of the first prevalence estimate, and $\mathrm{var}(\hat{p}_2)$ = the variance of the second prevalence estimate. In the first formula, $\mathrm{cov}(\hat{p}_1, \hat{p}_2)$ = 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 $\mathrm{SE}(\hat{p}_1)\mathrm{SE}(\hat{p}_2)$ = the product of the standard errors 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 standard errors). 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, like in the case of comparing two nonadjacent year estimates. Note that the first and second prevalence estimates may take the form of prevalence estimates from two different survey years (e.g., 2012 and 2013, respectively), prevalence estimates from sets of combined survey data (e.g., 2010–2011 annual averages and 2012–2013 annual averages, respectively), or prevalence estimates for 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 appropriate software.

Under the null hypothesis, the test statistic *t* is a random variable that asymptotically follows a *t*-distribution. Therefore, calculated values of *t*, along with the appropriate degrees of freedom, 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 1 above) will, in general, not be equal to 0. SUDAAN® is 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 (RTI International, 2012). A similar procedure and formula for *t* are used for estimated totals; however, it should be noted that because it was necessary to calculate the standard error outside SUDAAN for domains forced by the weighting process to match their respective U.S. Census Bureau population estimates, the corresponding test statistics also were computed outside SUDAAN. SAS®, SUDAAN, and Stata® examples showing the computational methods for generating *p* values of estimates of *t* and estimated totals can be found in Appendix A (Exhibits A.7 through A.18).

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 Section 10). As the degrees of freedom approach infinity, the *t*-distribution approaches the *Z* distribution. That is, because most of the statistical tests performed have 900 degrees of freedom, the *t* tests performed produce approximately the same numerical results as if a *Z* test had been performed.

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 were conducted first to control the error level for multiple comparisons. In Appendix A, see Exhibit A.27 for example SUDAAN code and Exhibit A.28 for example Stata 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, 2012).

If SUDAAN is not available to compute the significance testing, using published estimates can provide similar testing results. When comparing prevalence rates 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 and 7.2). 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 Example of Comparing Prevalence Estimates

The following example reproduces the difference in the proportions test between years 2012 and 2013 for a measure shown in Table 1.1B of the 2013 detailed tables (CBHSQ, 2014b). Table 1.1B displays the prevalence for lifetime, past year, and past month illicit drug use. This example will test the difference between 2012 and 2013 past year pain reliever use. Pain reliever use shown in Table 1.1B has a prevalence rate of 4.8 percent in 2012 and 4.2 percent in 2013. The corresponding standard errors shown in Table 1.1D are 0.14 percent for 2012 and 0.13 percent for 2013. Assuming the source data is not available and/or the user does not have access to appropriate software (i.e., SUDAAN), the second *t* test formula provided earlier in this section can be used with the assumption that the correlation is 0. Note, that

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

$$t_{900} = \frac{4.8 - 4.2}{\sqrt{0.14^2 + 0.13^2 - 2(0)(0.14)(0.13)}} = 3.1405 .$$

Using a *t* test to find the corresponding *p* value when $t = 3.1405$ and $df = 900$, results in *p* value = 0.0017. This is very close to the SUDAAN calculated *p* value of 0.0019 provided in Table 1.1P. This example confirms that the difference between the 2012 estimate of 4.8 percent and the 2013 estimate of 4.2 percent is statistically significant at the 0.01 level as indicated by footnote b included on the 2012 estimate in Table 1.1B. Note that the calculated *p* value assuming the correlation is 0 is smaller than the actual *p* value, which seems to contradict the earlier assertion that assuming the correlation is 0 results in a more conservative *p* value. However, this example produces a smaller *p* value due to the use of rounded estimates from the table (if the unrounded estimates had been available, the formula would yield a slightly larger *p* value than what is published in the tables). Below is an example using the same formula with the unrounded estimates and the covariance from SUDAAN. The extra digits and covariance change the *t*-score slightly, resulting in the published *p* value of 0.0019.

$$t_{900} = \frac{4.80254221 - 4.22345673}{\sqrt{(0.13668773)^2 + (0.13063372)^2 - 2(0.037909152)(0.13668773)(0.13063372)}} = 3.12245$$

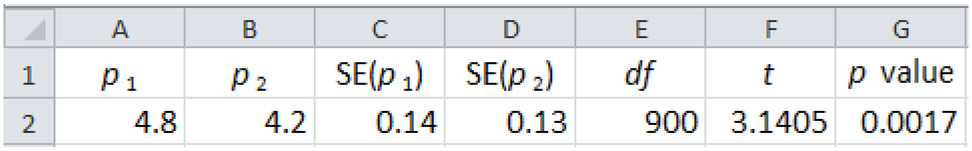
Also note that the correlations between estimates in adjacent years are generally very small and positive, but a negative correlation is possible. Estimates with negative correlations will also be close to 0; thus, the differences in SUDAAN-calculated *p* values and *p* values calculated from published estimates using the second *t* test formula provided earlier in this section (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.

7.2 Example of Comparing Prevalence Estimates in Excel

Using the same numbers presented in Section 7.1, 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 $\mathrm{var}(\hat{p}_1) = (\mathrm{SE}(\hat{p}_i))^2$. The correlation of 0 results in the simplified formula shown below (additionally the variances have been replaced by standard errors squared).

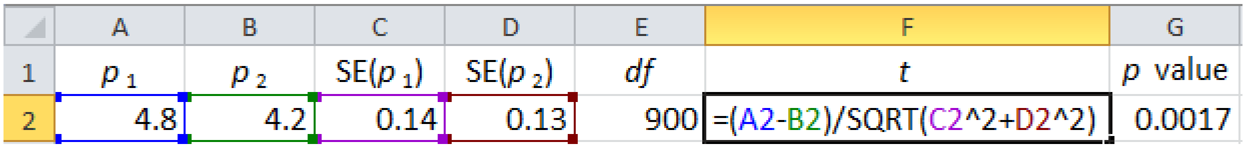
$$t_{df} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\mathrm{SE}(\hat{p}_1))^2 + (\mathrm{SE}(\hat{p}_2))^2}}$$

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, and then the formulas in columns F and G could be copied for all rows).



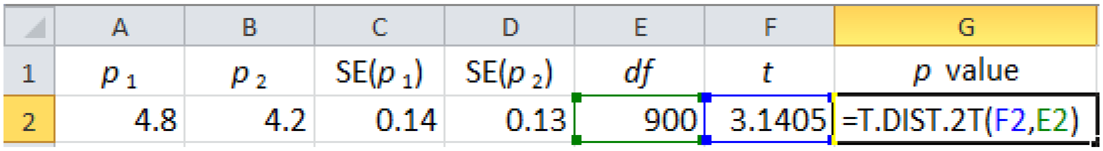
	A	B	C	D	E	F	G
1	*p* 1	*p* 2	SE(*p* 1)	SE(*p* 2)	*df*	*t*	*p* value
2	4.8	4.2	0.14	0.13	900	3.1405	0.0017

The standardized test statistic is found using the simplified formula for t_{df}.



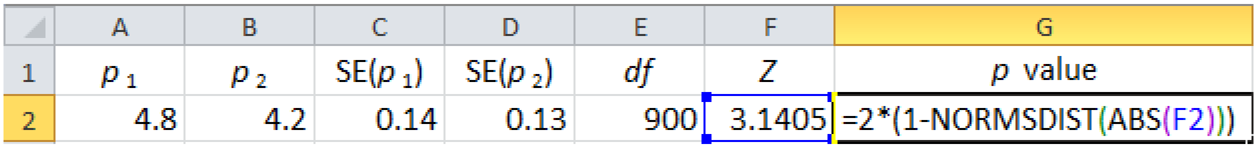
	A	B	C	D	E	F	G
1	*p* 1	*p* 2	SE(*p* 1)	SE(*p* 2)	*df*	*t*	*p* value
2	4.8	4.2	0.14	0.13	900	=(A2-B2)/SQRT(C2^2+D2^2)	0.0017

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	4.8	4.2	0.14	0.13	900	3.1405	=T.DIST.2T(F2,E2)

Alternatively, the Excel NORM.S.DIST function can be used to calculate the Standard Normal Cumulative Distribution Function since the *t*-distribution approaches the *Z* distribution as the degrees of freedom approach infinity. Tests performed having 900 degrees of freedom produce approximately the same numerical results as if a *Z* test had been performed. Note that this function refers to the test statistic as *Z* and does not require the degrees of freedom input.



	A	B	C	D	E	F	G
1	p_1	p_2	SE(p_1)	SE(p_2)	*df*	*Z*	*p* value
2	4.8	4.2	0.14	0.13	900	3.1405	=2*(1-NORMSDIST(ABS(F2)))

Both 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 rates shown with 95 percent CIs, it is important to note that two overlapping 95 percent CIs do not imply that their rates are statistically equivalent at the 5 percent level of significance. For additional information, see Schenker and Gentleman (2001) and Payton, Greenstone, and Schenker (2003).

This page intentionally left blank.

8. Confidence Intervals

In some National Survey on Drug Use and Health (NSDUH) publications, sampling error has been quantified using 95 percent confidence intervals (CIs). Frequently, NSDUH estimates are small percentages (i.e., are close to 0), and in that case, a logit transformation of the estimate provides favorable properties. For example, the logit transformation yields asymmetric interval boundaries between 0 and 1 that are more balanced with respect to the true probability that the true value falls below or above the interval boundaries. This is partly because for values close to 0, the distribution of a logit-transformed estimate approximates the normal distribution more closely than the standard estimate. Standard symmetric CIs for small proportions may also lead to the undesirable result of a lower CI limit that is less than 0.

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

$$L = 1\text{n}[P_d / (1 - P_d)],$$

where "1n" denotes the natural logarithm.

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

$$\hat{L} = 1\text{n}[\hat{p}_d /(1 - \hat{p}_d)].$$

The lower and upper confidence limits of *L* are formed as

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

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

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 standard error 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*). For example, to produce 95 percent confidence limits for national estimates, the value of *K* would be 1.96 based on 900 degrees of freedom (similarly, for large States, *K* would be 1.97 based on 192 degrees of freedom, and for small States, *K* would be 2.01 based on 48 degrees of freedom).

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 Section 10).

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

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

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

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

$$\hat{Y}_d = \hat{N}_d \cdot \hat{p}_d,$$

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 not fixed, the CI approximation assumes that the sampling variation in $\hat{N}_d$ is negligible relative to the sampling variation in $\hat{p}_d$.

Examples below illustrate how to compute and use confidence intervals. The example in Section 8.1 computes confidence intervals using the formulas shown above, the Section 8.2 example computes confidence intervals using Excel, the Section 8.3 example shows how to use the confidence intervals to compute standard errors, and the Section 8.4 example shows how to use Excel to compute the standard error from the confidence intervals.

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

The following example illustrates how to determine the 95 percent confidence interval using the prevalence estimates and standard errors provided for measures shown in the detailed tables and mental health detailed tables. This example will use estimates from Table 1.1B of the 2013 detailed tables (CBHSQ, 2014b), which displays the prevalence for lifetime, past year, and past month illicit drug use. This example will focus on 2013 past year pain reliever use. Pain reliever use shown in Table 1.1B has a prevalence rate of 4.2 percent in 2013. The corresponding standard error shown in Table 1.1D is 0.13 percent for 2013. This example uses the formulas shown above to determine the 95 percent confidence interval for the prevalence rate of past year pain reliever use in 2013. Note that

$$\text{var}(\hat{p}_d) = (\text{SE}(\hat{p}_d))^2;\ \text{thus,} \sqrt{\text{var}(\hat{p}_d)} = \text{SE}(\hat{p}_d).$$

Define log odds estimate:

$$\hat{L} = \ln[0.042/(1-0.042)] = -3.1281$$

Define the upper and lower confidence limits of the log odds:

$$A = -3.1281 - 1.96\left[\frac{0.0013}{0.0402}\right] = -3.1914$$

$$B = -3.1281 + 1.96\left[\frac{0.0013}{0.0402}\right] = -3.0648$$

Apply inverse logit transformation to yield confidence intervals p:

$$\hat{p}_{d,lower} = \frac{1}{1+\exp(3.1914)} = 0.0395$$

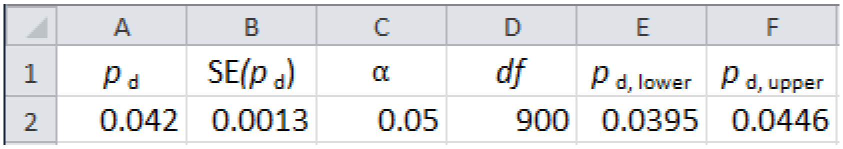
$$\hat{p}_{d,upper} = \frac{1}{1+\exp(3.0648)} = 0.0446$$

Rounding to two significant digits, the 95 percent confidence interval is 4.0 percent to 4.5 percent.

The same confidence interval calculated using SUDAAN® is also 4.0 percent to 4.5 percent, but note that the reduced number of significant digits shown in the published estimates may sometimes cause rounding errors when producing confidence intervals. However, the results are usually close. For examples using SUDAAN or Stata® to calculate confidence intervals, see Exhibits A.21 and A.22, respectively.

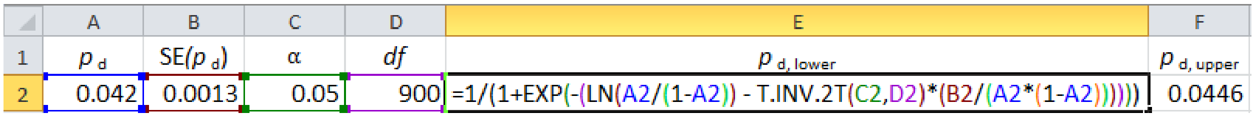
8.2 Example of Calculating Confidence Intervals in Excel Using Published Prevalence Estimates and Standard Errors

Using the same estimates presented in Section 8.1, this example uses Excel functions to produce the same confidence intervals produced in the previous example. Recall that $\text{var}(\hat{p}_d) = (\text{SE}(\hat{p}_d))^2$; thus, $\sqrt{\text{var}(\hat{p}_d)} = \text{SE}(\hat{p}_d)$. Excel can be used to set up a simple table (shown below) to produce the confidence interval. 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 confidence intervals (i.e., data for columns A through D would have to be entered for each row, and then the formulas in columns E and F could be copied for all rows).



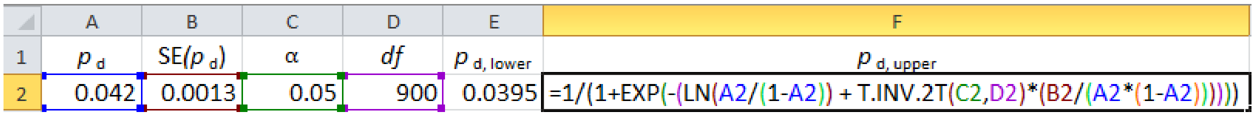
	A	B	C	D	E	F
1	p_d	SE(p_d)	α	*df*	$p_{d,\,lower}$	$p_{d,\,upper}$
2	0.042	0.0013	0.05	900	0.0395	0.0446

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)	α	*df*	*p* d, lower	*p* d, upper
2	0.042	0.0013	0.05	900	=1/(1+EXP(-(LN(A2/(1-A2)) - T.INV.2T(C2,D2)*(B2/(A2*(1-A2))))))	0.0446

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)	α	*df*	*p* d, lower	*p* d, upper
2	0.042	0.0013	0.05	900	0.0395	=1/(1+EXP(-(LN(A2/(1-A2)) + T.INV.2T(C2,D2)*(B2/(A2*(1-A2))))))

The 95 percent confidence interval is 4.0 percent to 4.5 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, degrees of freedom), where probability is the probability (between 0 and 1) for which you 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 confidence intervals, the alpha level is always 0.05. The example uses 900 degrees of freedom for a national estimate, 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 standard error for an estimate when only the prevalence and 95 percent confidence interval are provided. If a NSDUH publication provided only the prevalence rate for 2013 past year pain reliever use (4.2 percent) and the 95 percent confidence interval (4.0 percent to 4.5 percent), the reader may want to determine the standard error for use in significance testing. This example uses the formulas above to determine the standard error for the prevalence rate of past year pain reliever use in 2013. Note that

$$\mathrm{var}(\hat{p}_d) = (\mathrm{SE}(\hat{p}_d))^2;\ \text{thus, } \sqrt{\mathrm{var}(\hat{p}_d)} = \mathrm{SE}(\hat{p}_d).$$

Following is the formula to calculate A (lower confidence interval for log odds estimate) using the lower confidence interval of the prevalence rate (*p*).

$$\hat{p}_{d,lower} = \frac{1}{1+\exp(-A)};\ \text{thus, } A = \ln\left(\frac{\hat{p}_{d,lower}}{1-\hat{p}_{d,lower}}\right).$$

$$A = \ln\left(\frac{0.04}{1-0.04}\right) = 3.1781$$

Below is the formula for A (lower limit of the log odds ratio. To get the standard error, convert this formula as follows.

$$A = \hat{L} - K\left[\frac{\sqrt{\mathrm{var}(\hat{p}_d)}}{\hat{p}_d(1-\hat{p}_d)}\right]; \text{ thus, } \mathrm{SE}(\hat{p}_d) = \frac{(A-\hat{L})(\hat{p}_d(1-\hat{p}_d))}{-K}.$$

Recall from the Section 8.1 example that $\hat{L} = 3.1281$.Thus, the standard error is computed as follows:

$$SE(\hat{p}_d) = \frac{(-3.1781 + 3.1281)(0.042(1-0.042))}{-1.96} = 0.0010 \text{ or } 0.10\%$$

Using similar steps, the standard error can be produced from the upper confidence interval with the formulas below. Note that the denominator is positive in the standard error formula when using the upper confidence interval.

$$B = \ln\left(\frac{\hat{p}_{d,upper}}{1-\hat{p}_{d,upper}}\right) \text{ and } \mathrm{SE}(\hat{p}_d) = \frac{(B-\hat{L})(\hat{p}_d(1-\hat{p}_d))}{K}$$

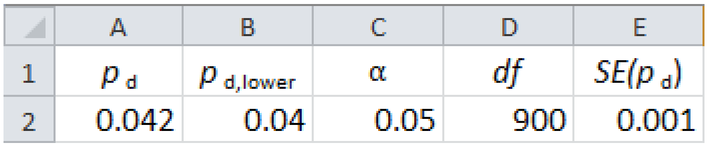
$$B = -3.0551 \text{ and } \mathrm{SE}(\hat{p}_d) = 0.0015, \text{ or } 0.15 \text{ percent}$$

As previously mentioned, Table 1.1D shows that the actual standard error when calculated in SUDAAN is 0.13 percent, which is close to both 0.10 percent and 0.15 percent. Note that the reduced number of significant digits shown in the published estimates may cause rounding errors when producing standard errors from the lower or upper limits of the confidence intervals. This can result in standard error estimates that differ when compared with the SUDAAN calculated standard error. However, standard errors calculated from the lower or upper limits usually will provide the same testing results as tests performed in SUDAAN, except results may differ when the *p* value is close to the predetermined level of significance.

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 standard errors from the previous example (i.e., the SUDAAN-generated standard error from Table 1.1D). Recall that $\mathrm{var}(\hat{p}_d) = (\mathrm{SE}(\hat{p}_d))^2$; thus, $\sqrt{\mathrm{var}(\hat{p}_d)} = \mathrm{SE}(\hat{p}_d)$. Excel can be used to set up a simple table (shown below) to produce the standard error from the upper and lower limits of the confidence interval. Cells A2 through D2 are the known values input by the user. Cell E2 contains the function to determine the standard error. This table could extend over several rows to aid in producing many standard errors (i.e., data for columns A through D would have to be entered for each row, and then the formula in column E could be copied for all rows). Note that once the methods used in this example have determined the standard error from the confidence interval, the methods shown in the Section 7.2 example can be used to perform independent *t* tests for differences of reported estimates in Excel.

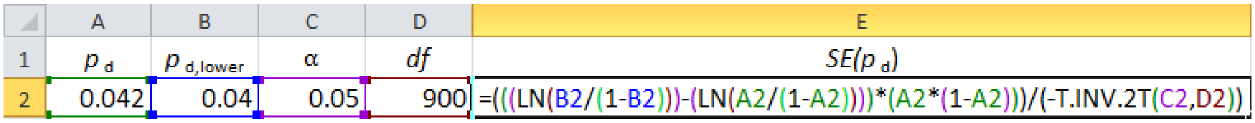
Calculate the standard error from the lower limit of the confidence interval:



	A	B	C	D	E
1	*p* d	*p* d,lower	α	*df*	*SE(p* d*)*
2	0.042	0.04	0.05	900	0.001

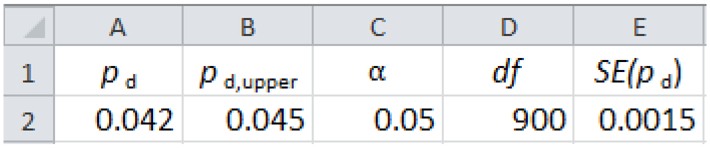
$\mathrm{SE}(\hat{p}_d) = 0.0010$, or 0.10 percent

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



	A	B	C	D	E
1	*p* d	*p* d,lower	α	*df*	*SE(p* d*)*
2	0.042	0.04	0.05	900	=(((LN(B2/(1-B2))-(LN(A2/(1-A2))))*(A2*(1-A2))/(-T.INV.2T(C2,D2))

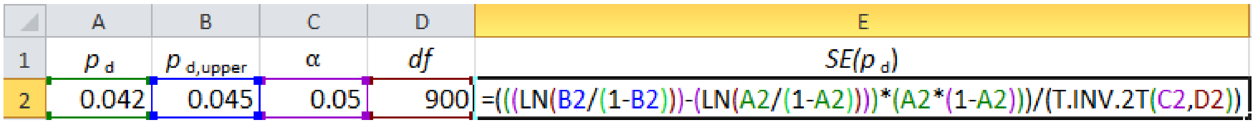
Calculate the standard error from the upper limit of the confidence interval:



	A	B	C	D	E
1	*p* d	*p* d,upper	α	*df*	*SE(p* d*)*
2	0.042	0.045	0.05	900	0.0015

$\mathrm{SE}(\hat{p}_d) = 0.0015$, or 0.15 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	α	*df*	*SE(p* d*)*
2	0.042	0.045	0.05	900	=(((LN(B2/(1-B2))-(LN(A2/(1-A2))))*(A2*(1-A2))/(T.INV.2T(C2,D2))

Remember that the reduced number of significant digits shown in the published estimates may cause rounding errors when producing standard errors. This can result in standard error estimates that differ when using the lower or upper limit when compared with the SUDAAN calculated standard error. However, standard errors calculated from the lower or upper limits usually will provide the same testing results as tests performed in SUDAAN, except results may differ when the *p* value is close to the predetermined level of significance.

9. Incidence Estimates

In epidemiological studies, incidence is defined as the number of new cases of a disease occurring within a specific period of time. Similarly, in substance use studies, incidence refers to the first use of a particular substance.

Starting with the 2004 National Survey on Drug Use and Health (NSDUH) data, the evaluation of trends in the initiation of drug use was presented by estimates of past year drug use incidence or initiation (i.e., the number of users whose first use was within the 12 months before their interview date). This incidence measure, termed "past year initiation," is determined by self-reported past year use, age at first use, year and month of most recent new use, and the interview date.

Since 1999, the NSDUH questionnaire allowed for the collection of year and month of first use for recent initiates (i.e., people who used a particular substance for the first time in a given survey year). Month, day, and year of birth also were obtained directly or imputed for item nonrespondents as part of the data processing. In addition, the questionnaire call record provided the date of the interview. By imputing a day of first use within the year and month of first use, a specific date of first use, $\text{(MM/DD/YYYY)}_{\text{First Use of Substance}}$, can be used for estimation purposes.

Past year initiation among people using a substance in the past year can be viewed as an indicator variable defined as follows:

$$I_{\text{(Past Year Initiate)}} \text{ if } [\text{(MM/DD/YYYY)}_{\text{Interview}} - \text{(MM/DD/YYYY)}_{\text{First Use of Substance}}] \le 365,$$

where $\text{(MM/DD/YYYY)}_{\text{Interview}}$ denotes the month, day, and year of the interview, and $\text{(MM/DD/YYYY)}_{\text{First Use of Substance}}$ denotes the date of first use.

Note that the 12-month reference period (i.e., 365 days) is set up on the calendar at the beginning of the audio computer-assisted self-interviewing portion of the computer-assisted interview. For example, if the date of the interview (DOI) is December 1, 2013 (12/01/2013), then 365 days earlier would be December 1, 2012 (12/01/2012). If a respondent's date of first use is the same as the DOI, then the respondent is considered a past year initiate (since I = 0). Additionally, in this example, a respondent interviewed on 12/01/2013 could have used for the first time as far back as 12/01/2012 and be considered a past year initiate.

The calculation of past year initiation does not take into account whether the respondent initiated substance use while a resident of the United States. This method of calculation has little effect on past year estimates and provides direct comparability with other standard measures of substance use because the populations of interest for the measures will be the same (i.e., both measures examine all possible respondents and do not restrict to those only initiating substance use in the United States).

One important note for incidence estimates is the relationship between a main substance category and subcategories of substances (e.g., illicit drugs would be a main category and inhalants and marijuana would be examples of subcategories in relation to illicit drugs). For most measures of substance use, any member of a subcategory is by necessity a member of the main category (e.g., if a respondent is a past month user of a particular drug, then he or she is also a past month user of illicit drugs in general). However, this is not the case with regard to incidence statistics. Because an individual can only be an initiate of a particular substance category (main or sub) a single time, a respondent with lifetime use of a subcategory may not, by necessity, be included as an initiate of the corresponding main category, even if he or she were an initiate for a different subcategory.

In addition to estimates of the number of people initiating use of a substance in the past year, estimates of the mean age of past year first-time users of these substances were computed. Unless specified otherwise, estimates of the mean age at initiation in the past 12 months have been restricted to people aged 12 to 49 so that the mean age estimates reported are not influenced by those few respondents who were past year initiates at age 50 or older. As a measure of central tendency, means are influenced heavily by the presence of extreme values in the data, and this constraint should increase the utility of these results to health researchers and analysts by providing a better picture of the substance use initiation behaviors among the civilian, noninstitutionalized population in the United States. This constraint was applied only to estimates of mean age at first use and does not affect estimates of incidence.

Because NSDUH is a survey of people aged 12 or older at the time of the interview, younger individuals in the sample dwelling units are not eligible for selection into the NSDUH sample. Some of these younger people may have initiated substance use during the past year. As a result, past year initiate estimates suffer from undercoverage when one can think of the estimates as reflecting all initial users regardless of current age. For earlier years, data can be obtained retrospectively based on the age at and date of first use. As an example, people who were 12 years old on the date of their interview in the 2013 survey may have reported initiating use of cigarettes between 1 and 2 years ago; these people would have been past year initiates reported in the 2012 survey had people who were 11 years old on the date of the 2012 interview been allowed to participate in the survey. Similarly, estimates of past year use by younger people (aged 10 or younger) can be derived from the current survey, but they apply to initiation in prior years—not the survey year.

To get an impression of the potential undercoverage in the current year, reports of substance use initiation reported in 2013 by people aged 12 or older were estimated for the years in which these people would have been 1 to 11 years younger. These estimates do not necessarily reflect behavior by people who were 1 to 11 years younger in 2013. Instead, the data for the 11-year-olds reflect initiation in the year before the 2013 survey, the data for the 10-year-olds reflect behavior between the 12th and 23rd month before the 2013 survey, and so on. A rough way to adjust for the difference in the years that the estimate pertains to without considering changes to the population is to apply an adjustment factor to each age-based estimate of past year initiates. The adjustment factor can be based on a ratio of lifetime users aged 12 to 17 in 2013 to the same estimates for the prior applicable survey year. To illustrate the calculation, consider past year use of alcohol. In the 2013 survey, 101,441 people who were 12 years old were estimated to have initiated use of alcohol between 1 and 2 years earlier. These people would have been past year initiates in the 2012 survey conducted on the same dates had the 2012 survey covered younger people. The estimated number of lifetime users currently aged 12 to 17 was 7,669,220 for 2013 and 8,067,487 for 2012, indicating fewer overall initiates of alcohol use among people aged 17 or younger in 2013. Thus, an adjusted estimate of initiation of alcohol use by people who were 11 years old in 2013 is given by

$$(\text{Estimated Past Year Initates Aged 11})_{2012} \times \frac{(\text{Estimated Lifetime Users Aged 12 to 17})_{2013}}{(\text{Estimated Lifetime Users Aged 12 to 17})_{2012}} \ .$$

This yielded an adjusted estimate of 96,433 people who were 11 years old on a 2013 survey date and initiated use of alcohol in the past year:

$$101{,}441 \times \frac{7{,}669{,}220}{8{,}067{,}487} = 96{,}433$$

A similar procedure was used to adjust the estimated number of past year initiates among people who would have been 10 years old on the date of the interview in 2011 and for younger people in earlier years. The overall adjusted estimate for past year initiates of alcohol use by people aged 11 or younger on the date of the interview was 161,183, or about 3.5 percent of the estimate based on past year initiation by people aged 12 or older only (161,183 ÷ 4,558,527 = 0.0354). Based on similar analyses, the estimated undercoverage of past year initiates was 2.3 percent for cigarettes, 1.1 percent for marijuana, and 13.4 percent for inhalants.

The undercoverage of past year initiates aged 11 or younger also affects the mean age-at-first-use estimate. An adjusted estimate of the mean age at first use was calculated using a weighted estimate of the mean age at first use based on the current survey and the numbers of people aged 11 or younger in the past year obtained in the aforementioned analysis for estimating undercoverage of past year initiates. Analysis results showed that the mean age at first use was changed from 17.3 to 17.0 for alcohol, from 17.8 to 17.6 for cigarettes, from 18.0 to 17.9 for marijuana, and from 19.2 to 17.7 for inhalants. The decreases reported above are comparable with results generated in prior survey years.

This page intentionally left blank.

10. Suppression of Estimates with Low Precision

Direct survey estimates that were considered to be unreliable because of unacceptably large sampling errors were not reported, but rather were noted by an asterisk (*). The criteria used to assess the need to suppress direct survey estimates were based on prevalence (for proportion estimates), the relative standard error (RSE) (defined as the ratio of the standard error [SE] over the estimate), nominal (actual) sample size, and effective sample size for each estimate.

Proportion estimates ($\hat{p}$), or rates, within the range $0 < \hat{p} < 1$, and corresponding estimated numbers of users were suppressed if

$$\text{RSE}[-\ln(\hat{p})] > .175 \text{ when } \hat{p} \le .5$$

or

$$\text{RSE}[-\ln(1-\hat{p})] > .175 \text{ when } \hat{p} > .5 \,.$$

The choice of .175 is arbitrary, but it roughly marks the tails of the distribution.

Based on a first-order Taylor series approximation of $\text{RSE}[-\ln(\hat{p})]$ and $\text{RSE}[-\ln(1-\hat{p})]$, the following equation was derived and used for computational purposes when applying a suppression rule dependent on effective sample sizes:

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

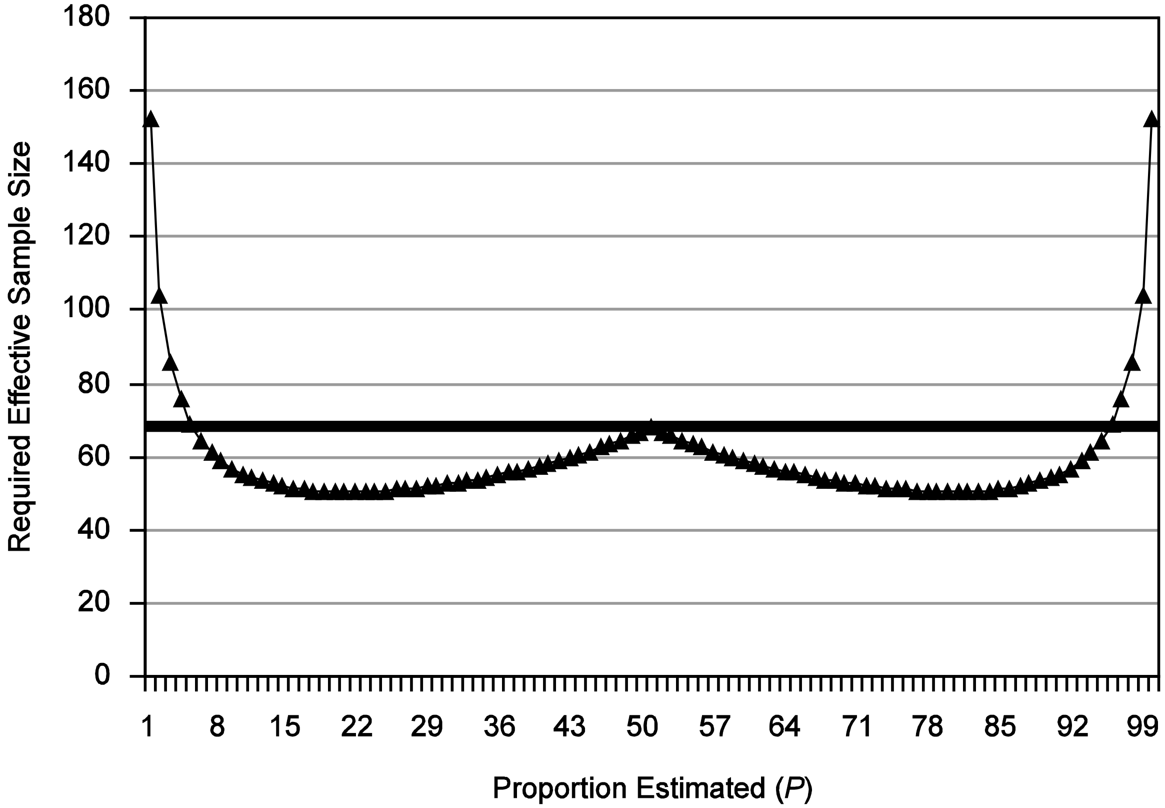
or

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

The separate formulas for $\hat{p} \le .5$ and $\hat{p} > .5$ produce a symmetric suppression rule; that is, if $\hat{p}$ is suppressed, $1-\hat{p}$ will be suppressed as well. See Figure 10.1 for a graphical representation of the required minimum effective sample sizes as a function of the proportion estimated. When $.05 < \hat{p} < .95$, the symmetric properties of the rule produce local minimum effective sample sizes at $\hat{p}$ = .2 and again at $\hat{p}$ = .8, such that an effective sample size of greater than 50 is required; this means that estimates would be suppressed for these values of $\hat{p}$ unless the effective sample sizes were greater than 50. Within this same interval of $.05 < \hat{p} < .95$, a local maximum effective sample size of 68 is required at $\hat{p}$ = .5. So, to simplify requirements and maintain a conservative suppression rule, estimates of $\hat{p}$ between .05 and .95, which had effective sample sizes below 68, were suppressed.

The effective sample size for a domain is a function of the nominal sample size and the design effect (i.e., nominal sample size/design effect). During the original development of this suppression rule, the design effect was calculated outside SUDAAN® (RTI International, 2012) in SAS®. Since the 2005 National Survey on Drug Use and Health (NSDUH) analysis, the direct SUDAAN design effect was used to provide a more precise and accurate reflection of the design effect (due to the removal of several possible rounding errors) when compared with the SAS method used in the past. The differences between the direct SUDAAN design effects and the SAS-calculated design effects occur only at approximately the tenth decimal place or later; however, previously published estimates that were on the borderline of being suppressed or unsuppressed due to the effective sample size suppression rule may potentially change from suppressed to unsuppressed, or vice versa.

Figure 10.1 Required Effective Sample in the 2013 NSDUH as a Function of the Proportion Estimated



In addition, a minimum nominal sample size suppression criterion ($n = 100$) that protects against unreliable estimates caused by small design effects and small nominal sample sizes was employed. Table 10.1 shows a formula for calculating design effects. Prevalence estimates also were suppressed if they were close to 0 or 100 percent (i.e., if $\hat{p} < .00005$ or if $\hat{p} \geq .99995$).

Beginning with the 1991 survey, the suppression rule for proportions based on $\text{RSE}[-1\text{n}(\hat{p})]$ described above replaced an older rule in which data were suppressed whenever $\text{RSE}(\hat{p}) > .5$. This rule was changed because the older rule imposed a very stringent application for small $\hat{p}$, but a very lax application for large $\hat{p}$. The new rule ensured a more uniformly stringent application across the whole range of $\hat{p}$ (i.e., from 0 to 1). The old rule also was asymmetric in the sense that suppression only occurred in terms of $\hat{p}$; that is, there was no complementary rule for $(1 - \hat{p})$, which the new suppression rules now account for.

Estimates of totals were suppressed if the corresponding prevalence rates were suppressed. Estimates of means not bounded between 0 and 1 (e.g., mean age at first use, mean number of drinks consumed) were suppressed if the RSEs of the estimates were larger than .5 or if the sample sizes were 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 appears to provide sufficient precision and still allow reporting by year of first use for many substances. In these cases, the totals (e.g., total number of drinks consumed) were suppressed if the corresponding mean estimates were suppressed.

Section 4 of the detailed tables demonstrates an exception to the rule that indicates the totals are suppressed when their corresponding means are suppressed. Some tables in Section 4 of the detailed tables show estimates of incidence among different populations. Specifically, these Section 4 tables display the number of initiates among three different populations: the total population, people at risk for initiation, and past year users. In these tables, some mean estimates may be suppressed while the total estimate is not suppressed. When at least one mean estimate in the table is not suppressed, one can assume that the numerator (or total estimate) is not the cause for the suppression and the total estimate will not be suppressed. In contrast, when all mean estimates are suppressed, the total will also be suppressed.

Tables that show sample sizes and population counts do not incorporate the suppression rule for several reasons. One reason is that no mean is associated with these 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.

The suppression criteria for various NSDUH estimates are summarized in Table 10.1, and sample SAS and Stata® code demonstrating how to implement these rules can be found in Appendix A (Exhibits A.5 and A.6).

Table 10.1 Summary of 2013 NSDUH Suppression Rules

Estimate	**Suppress if:**
Prevalence Rate, $\hat{p}$, with Nominal Sample Size, *n*, and Design Effect, *deff* $\left(deff = \frac{n[\mathrm{SE}(\hat{p})]^2}{\hat{p}(1-\hat{p})} \right)$	(1) The estimated prevalence rate, $\hat{p}$, is < 0.00005 or ≥ 0.99995, or (2) $\frac{\mathrm{SE}(\hat{p})\ /\ \hat{p}}{-\ln(\hat{p})} > 0.175$ when $\hat{p} \leq 0.5$, or $\frac{\mathrm{SE}(\hat{p})\ /\ (1-\hat{p})}{-1\mathrm{n}(1-\hat{p})} > .175$ when $\hat{p} > 0.5$, or (3) Effective *n* < 68, where $Effective\ n = \frac{n}{deff} = \frac{\hat{p}(1-\hat{p})}{\left[\mathrm{SE}(\hat{p})\right]^2}$, or (4) *n* < 100. Note: The rounding portion of this suppression rule for prevalence rates will produce some estimates that round at one decimal place to 0.0 or 100.0 percent but are not suppressed from the tables.
Estimated Number (Numerator of $\hat{p}$)	The estimated prevalence rate, $\hat{p}$, is suppressed. Note: In some instances when $\hat{p}$ is not suppressed, the estimated number may appear as a 0 in the tables. This means that the estimate is greater than 0 but less than 500 (estimated numbers are shown in thousands). Note: In some instances when totals corresponding to several different means that are displayed in the same table and some, but not all, of those means are suppressed, the totals will not be suppressed. When all means are suppressed, the totals will also be suppressed.
Means not bounded between 0 and 1 (i.e., Mean Age at First Use, Mean Number of Drinks), $\bar{x}$, with Nominal Sample Size, *n*	(1) $\mathrm{RSE}(\bar{x}) > 0.5$, or (2) *n* < 10.

deff = design effect; RSE = relative standard error; SE = standard error.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2013.

References

Aldworth, J., Kott, P., Yu, F., Mosquin, P., & Barnett-Walker, K. (2012). Analysis of effects of 2008 NSDUH questionnaire changes: Methods to adjust adult MDE and SPD estimates and to estimate SMI in the 2005-2009 surveys. In *2010 National Survey on Drug Use and Health: Methodological resource book* (Section 16b, prepared for the Substance Abuse and Mental Health Services Administration under Contract No. HHSS283200800004C, Deliverable No. 39, RTI/0211838.108.005). Research Triangle Park, NC: RTI International.

American Psychiatric Association. (2008). *Diagnostic and statistical manual of mental disorders, 4th ed., text revision (DSM-IV-TR)*. Retrieved from http://www.psychiatry.org/practice/dsm/

Center for Behavioral Health Statistics and Quality. (2010). *Results from the 2009 National Survey on Drug Use and Health: Mental health detailed tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2012a). *Results from the 2010 National Survey on Drug Use and Health: Mental health detailed tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2012b). *Results from the 2011 National Survey on Drug Use and Health: Detailed tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2012c). *Results from the 2011 National Survey on Drug Use and Health: Mental health detailed tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2012d). *Results from the 2011 National Survey on Drug Use and Health: Mental health findings* (HHS Publication No. SMA 12-4725, NSDUH Series H-45). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2012e). *Results from the 2011 National Survey on Drug Use and Health: Summary of national findings* (HHS Publication No. SMA 12-4713, NSDUH Series H-44). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2013). *Results from the 2012 National Survey on Drug Use and Health: Mental health detailed tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2014a). 2012 Mental Health Surveillance Study: Design and estimation report. In *2012 National Survey on Drug Use and Health: Methodological resource book (Section 16a)*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Center for Behavioral Health Statistics and Quality. (2014b). *Results from the 2013 National Survey on Drug Use and Health: Detailed tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2014c). *Results from the 2013 National Survey on Drug Use and Health: Mental health detailed tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2014d). *Results from the 2013 National Survey on Drug Use and Health: Mental health findings* (HHS Publication No. SMA 14- 4887, NSDUH Series H-49). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2014e). *Results from the 2013 National Survey on Drug Use and Health: Summary of national findings* (HHS Publication No. SMA 14-4863, NSDUH Series H-48). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Center for Behavioral Health Statistics and Quality. (2014f). Sample design report. In *2013 National Survey on Drug Use and Health: Methodological resource book (Section 2)*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Center for Behavioral Health Statistics and Quality. (2015a). Editing and imputation report. In *2013 National Survey on Drug Use and Health: Methodological resource book (Section 10)*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Center for Behavioral Health Statistics and Quality. (2015b). Person-level sampling weight calibration. In *2013 National Survey on Drug Use and Health: Methodological resource book (Section 11)*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. (2012). *National Survey on Drug Use and Health: 2010 public use file codebook*. Retrieved from http://www.icpsr.umich.edu/icpsrweb/SAMHDA/studies/32722

Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. (2014). *National Survey on Drug Use and Health: 2013 public use file and codebook*. Retrieved from http://www.icpsr.umich.edu/icpsrweb/SAMHDA/studies/35509

Chen, P., Cribb, D., Dai, L., Gordek, H., Laufenberg, J., Sathe, N., & Westlake, M. (2013). Person-level sampling weight calibration. In *2011 National Survey on Drug Use and Health: Methodological resource book* (Section 12, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. HHSS283200800004C, Phase II, Deliverable No. 39, RTI/0211838.207.004). Research Triangle Park, NC: RTI International.

Chromy, J. R., & Penne, M. (2002). Pair sampling in household surveys. In *Proceedings of the 2002 Joint Statistical Meetings, American Statistical Association, Survey Research Methods Section, New York, NY [CD-ROM]* (pp. 552-554). Alexandria, VA: American Statistical Association. [Available as a PDF at http://www.amstat.org/sections/SRMS/Proceedings/]

Dean, E., & LeBaron, P. (2009, November). *2008 National Survey on Drug Use and Health: Context effects report* (prepared for the Substance Abuse and Mental Health Services Administration under Contract No. 283-2004-00022, RTI/0209009.523.006.002). Research Triangle Park, NC: RTI International.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (2002, November). *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP)*. New York, NY: New York State Psychiatric Institute, Biometrics Research.

Hughes, A., Muhuri, P., Sathe, N., & Spagnola, K. (2012). *State estimates of substance use and mental disorders from the 2009-2010 National Surveys on Drug Use and Health* (HHS Publication No. SMA 12-4703, NSDUH Series H-43). Rockville, MD: Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. [Available at http://samhsa.gov/data/]

Office of Applied Studies. (2009a). *Results from the 2008 National Survey on Drug Use and Health: Detailed tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

Office of Applied Studies. (2009b). *Results from the 2008 National Survey on Drug Use and Health: National findings* (HHS Publication No. SMA 09-4434, NSDUH Series H-36). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://samhsa.gov/data/]

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.

Rehm, J., Üstün, T. B., Saxena, S., Nelson, C. B., Chatterji, S., Ivis, F., & Adlaf, E. (1999). On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. *International Journal of Methods in Psychiatric Research, 8*(2), 110-123.

RTI International. (2012). *SUDAAN® language manual, Release 11.0.0*. Research Triangle Park, NC: RTI International.

Ruppenkamp, J., Emrich, S., Aldworth, J., Hirsch, E., & Foster, M. (2006, February). *Missingness evaluation in the 2004 NSDUH* (draft report, prepared for the Substance Abuse and Mental Health Services Administration under Contract No. 283-03-9028, RTI/0208726.187.022). Research Triangle Park, NC: RTI International.

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.

This page intentionally left blank.

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

This appendix provides guidance concerning various options that should be specified in both SUDAAN® and Stata® to correctly analyze the National Survey on Drug Use and Health (NSDUH) data. Additionally, example SAS®, SUDAAN® Software for Statistical Analysis of Correlated Data (RTI International, 2012), and Stata code is provided to illustrate how the information in this report is applied to generate estimates (means, totals, and percentages along with the standard errors [SEs]), implement the suppression rule, perform statistical tests of differences, handle missing data, calculate confidence intervals, test between overlapping domains, and test independence of two variables. Specifically, the examples produce estimates of past month alcohol use by year (2012 and 2013) and gender (males and females) using the statistical procedures documented within this report and implemented in the 2013 detailed tables (Center for Behavioral Health Statistics and Quality [CBHSQ], 2014b) and the 2013 mental health detailed tables (CBHSQ, 2014c). The examples below are created using variable names found on the restricted-use dataset; thus, some variable names may differ when using the public use file (see footnote 2 for more detail). Note that all the detailed tables and mental health detailed tables are produced using SAS and SUDAAN code. However, the Stata code below replicates results from these tables. The exhibit number for each example, a description of the example, and a reference to the report section that addresses the example are provided in Table A.1.

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

SAS/SUDAAN Exhibit	Stata Exhibit	Description	Report Section
A.1	A.2	Produces estimates (including means, totals, and the respective standard errors).	Section 3
A.3	A.4	Calculates the standard error of the total for controlled domains using the estimates produced in Exhibits A.1 and A.2.	Section 5
A.5	A.6	Creates suppression indicators for each estimate (i.e., suppression rule).	Section 10
A.7	A.8	Performs statistical tests of differences between means.	Section 7
A.9	A.10	Calculates the *p* value for the test of differences between uncontrolled totals (using estimates produced in Exhibits A.7 and A.8).	Section 7
A.11, A.13, A.15, and A.17	A.12, A.14, A.16, and A.18	Calculates the *p* value for the test of differences between controlled domains by producing the covariance matrix, pulling the relevant covariance components, and calculating the variances.	Section 7
A.19	A.20	Produces estimates where the variable of interest has missing values.	Section 4
A.21	A.22	Calculates a confidence interval using estimates produced in Exhibits A.1 and A.2.	Section 8
A.23	A.24	Calculates percentages and the associated standard errors.	Section 3
A.25	A.26	Performs statistical tests of differences between two groups when the two groups overlap.	Section 7
A.27	A.28	Performs tests of the independence of the prevalence variable and subgroup variable.	Section 7

Guide for Defining Options for Analyzing NSDUH Data

Before running the SUDAAN procedures, the input dataset must be sorted by the nesting variables (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. The SUDAAN procedure DESCRIPT can then be run to produce weighted and unweighted sample sizes, means, totals, SEs of means and totals, and *p* values for testing of the means and totals.

Stata commands can be run without the data being sorted. The Stata commands svy: mean and svy: total will be used throughout in these exhibits (note that Stata still uses VESTR and VEREP but the data do not need to be sorted).

The following options are specified within the SUDAAN and Stata examples to correctly produce estimates using NSDUH data.

DESIGN

Due to the NSDUH sample design, estimates are calculated using a method in SUDAAN that is unbiased for linear statistics. This method is based on multistage clustered sample designs where the first-stage (primary) sampling units are drawn with replacement. In SUDAAN, a user must specify DESIGN=WR (meaning with replacement). Note that with Stata the design does not need to be indicated, because the svyset command uses Taylor linearized variance estimation as a default.

Nesting Variables (VESTR and VEREP)

The nesting variables are used to capture explicit stratification and to identify clustering with the NSDUH data, which are needed to compute the variance estimates correctly. 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, thus constituting the 50 percent overlap between survey years. 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 be listed first, followed by the VEREP variable. In Stata, the nesting variables are specified in the svyset command. Unlike the svyset command in Stata, the NEST statement will need to be used each time a user calls one of the appropriate SUDAAN procedures.

Degrees of Freedom (DDF or DOF)

As described in Section 6 of this report, the degrees of freedom (*df*) are 900 for national estimates, 192 for large States (California, Florida, Illinois, Michigan, New York, Pennsylvania, Ohio, Texas), and 48 for all other States. For an analysis of a group of States, the degrees of freedom can be less than or equal to the sum of the degrees of freedom for each individual State due to overlap of variance strata. 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 can also be used for analyses combining survey data across years. When combining any years of data from 2005 through 2012, the degrees of freedom remain the same as if it were a single year (e.g., 900 for national estimates) because these years are part of the same sample design. When comparing estimates in two domains with different degrees of freedom, err on the conservative side and use the smaller degrees of freedom. 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 (i.e., "dof(900)"). If switching from national estimates to State estimates, the svyset command would need to be rerun with the updated degrees of freedom.

Design Effect

The option DEFT4 within SUDAAN provides the correct measure of variance inflation due to stratification (or blocking), clustering, and unequal weighting in NSDUH estimation. Requesting deff srssubpop in Stata gives the same result as using DEFT4 in SUDAAN.

The following SAS, SUDAAN, and Stata examples apply the specific NSDUH options described previously to compute estimates, apply the suppression rule, and perform significance testing by using the data produced by the examples in Exhibit A.1 (using SUDAAN code) and Exhibit A.2 (using Stata code).

Generation of Estimates

Exhibits A.1 and A.2 demonstrate how to compute various types of estimates for past month alcohol use by year and gender using the SUDAAN descript procedure and the Stata svy: mean and svy: total commands, respectively. The SUDAAN example includes code 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 totals (SETOTAL). The Stata svy: mean and svy: total commands will produce the same estimates. Whether the SETOTAL is taken directly from SUDAAN or Stata depends on whether the specified domain (i.e., gender in this example) is among those forced to match their respective U.S. Census Bureau population estimates through the weight calibration process. See the Standard Errors section below for additional information.

Exhibit A.1 SUDAAN DESCRIPT Procedure (Estimate Generation)

```
PROC SORT DATA=DATANAME; /*SAS code to sort output dataset by
Nesting Variables*/
BY VESTR VEREP;
RUN;
```

Exhibit A.1 SUDAAN DESCRIPT Procedure (Estimate Generation) (continued)

```
      PROC DESCRIPT DATA=DATANAME DDF=900 DESIGN=WR FILETYPE=SAS DEFT4;
      NEST VESTR VEREP;
      WEIGHT ANALWT;   /*Standard single-year, person-level analysis
      weight*/

      VAR ALCMON;   /*Past month alcohol analysis variable*/
      SUBGROUP YEAR IRSEX;
           /*Year variable, where 2012=1 & 2013=2*/
           /*Gender variable, where male=1 & female=2*/
      LEVELS 2 2;
      TABLES YEAR*IRSEX; /*Gender 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 GENDER";
      RUN;

Note: The following CLASS statement could be used in place of SUBGROUP
and LEVELS statements in the above example:
      CLASS YEAR IRSEX;
```

Exhibit A.2 Stata COMMANDS svy: mean and svy: total (Estimate Generation)

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

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

/*ID Nesting variables (VESTR and VEREP) and weight variable (ANALWT -
standard single-year, person-level analysis weight*/
svyset verep [pw=analwt], strata(vestr) dof(900)

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

/*Estimated means of past month alcohol use by year and gender*/

  /*Year variable, where 2012=1 & 2013=2*/
  /*Gender variable, where male=1 & female=2*/
svy: mean alcmon, over(year irsex)
```

Exhibit A.2 Stata COMMANDS svy: mean and svy: total (Estimate Generation) (continued)

```
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*/
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 year and gender.*/
  local counter=1
    forvalues i=1/2 { /*number of years*/
      forvalues j=1/2 { /* number of gender categories*/
       replace mean_out=(M[1,`counter']) if year==`i' & irsex==`j'
       replace semean=(sqrt(S[`counter',`counter'])) ///
if year==`i' & irsex==`j'
       replace nsum=(N[1,`counter']) if year==`i' & irsex==`j'
       replace wsum=(W[1,`counter']) if year==`i' & irsex==`j'
       replace deffmean=(D[1,`counter']) if year==`i' & irsex==`j'
  local counter=`counter'+1
      }
    }

/*Estimated Totals*/
svy: total alcmon, over(year 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 gender.*/

  local counter=1
    forvalues i=1/2 { /*number of years*/
      forvalues j=1/2 { /* number of gender categories*/
        replace total_out=(M[1,`counter']) if year==`i' & irsex==`j'
        replace setotal=(sqrt(S[`counter',`counter'])) ///
if year==`i' & irsex==`j'
   local counter=`counter'+1
      }
    }
```

Exhibit A.2 Stata COMMANDS svy: mean and svy: total (Estimate Generation) (continued)

```
keep wsum mean_out semean total_out setotal nsum deffmean year irsex

duplicates drop year 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 year and gender*/
list year irsex wsum nsum mean_out semean total_out setotal

/*The output from this exhibit will be utilized in Exhibit A.16. 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 ".\\EXa2.dta" , replace
```

Standard Errors

As discussed in Section 5 of this report, the SE for the mean (or proportion) comes directly out of SUDAAN in the output variable SEMEAN (Exhibit A.1), and the SEMEAN is calculated in Stata by taking the square root of the variance (Exhibit A.2). However, to compute the SE of the totals, NSDUH implements different methods depending on whether the specified domain (i.e., gender in this example) is controlled or uncontrolled through poststratification during the weighting process. If a domain is uncontrolled (i.e., it is not one of the domains described in Table 5.1 in Section 5), then the SE of the total comes directly out of SUDAAN in the output variable SETOTAL. If the domain is controlled (i.e., it is one of the domains described in Table 5.1), then the SE of the total is calculated as SETOTAL (SE of controlled domain) = WSUM (weighted sample size) × SEMEAN (SE for the mean/proportion). Because gender is controlled, the SE of the totals would not be taken directly from the examples in Exhibits A.1 and A.2 but rather would be computed using the formula shown in Exhibits A.3 and A.4 (note that the formula is the same in both exhibits) (Exhibits A.1 and A.3 using SUDAAN/SAS code and Exhibits A.2 and A.4 using Stata code).

Exhibit A.3 SAS Code (Calculation of Standard Error of Totals for Controlled Domains)

```
DATA ESTIMATE;
SET OUT.SUDFILE; /*input the output file from above SUDAAN
                  procedure*/
/**************************************************************
  Define SETOTAL for gender because it is a controlled domain.
   In the SUDAAN procedure in Exhibit A.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.
**************************************************************/

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

RUN;
```

Exhibit A.4 Stata Code (Calculation of Standard Error of Totals for Controlled Domains)

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

Suppression Rule

As described in Section 10 of the report, each published NSDUH estimate goes through a suppression rule to detect if the estimate is unreliable due to an unacceptably large sampling error. The suppression rules as they apply to different types of estimates are shown in Table 10.1 in Section 10. The examples in Exhibits A.5 (SAS code) and A.6 (Stata code) show both the prevalence rate 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 rate 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., incidence tables in Section 4 of the 2013 detailed tables [CBHSQ, 2014b]), suppression is not as straightforward as coding the rule in the SAS/SUDAAN or Stata programs. As discussed in Section 10, perhaps 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. The best way to ensure that this happens is to program the total estimate in the table to be suppressed if, and only if, the mean with the largest denominator is suppressed. The analyst should also check the final table to ensure that the suppression follows the rule after the program has been run.

Exhibit A.5 SAS Code (Implementation of Suppression Rule)

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

/******APPLY THE PREVALENCE RATE SUPRESSION RULE*******/

/* CALCULATE THE RELATIVE STANDARD ERROR */
     IF MEAN GT 0.0 THEN RSE=SEMEAN/MEAN;

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

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

/*SUPRESSION RULE FOR PREVALENCE RATES*/
IF (MEAN LT .00005) OR (MEAN GE 0.99995) OR (RSELNP GT 0.175) OR
(EFFNSUM < 68) OR (NSUM <100) THEN SUPRULE=1;

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

RUN;
```

Exhibit A.6 Stata Code (Implementation of Suppression Rule)

```
/******APPLY THE PREVALENCE RATE SUPRESSION RULE*******/

/*CALCULATE THE RELATIVE STANDARD ERROR*/
generate rse=.
replace rse=semean/mean_out ///
if mean_out > 0.0 & !missing(mean_out)

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

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

Exhibit A.6 Stata Code (Implementation of Suppression Rule) (continued)

```
/*SUPRESSION RULE FOR PREVALENCE RATES*/
generate suprule1a=1 if rselnp > 0.175 & !missing(rselnp)
generate suprule1b=1 if mean_out < .00005 & !missing(mean)
generate suprule1c=1 if mean_out >= .99995 & !missing(mean)
generate suprule2=1 if effnsum < 68 & !missing(nsum)
generate suprule3=1 if nsum < 100 & !missing(nsum)

generate supress=0
replace supress=1 if suprule1a==1 | suprule1b==1 | ///
suprule1c==1 | suprule2==1 | suprule3==1

/*SUPRESSION RULE FOR MEANS NOT BOUNDED BY 0 AND 1, I.E. AVERAGES
(COMMENTED OUT FOR THIS EXAMPLE)*/
/*generate suprule=1 if (nsum < 100 & !missing(nsum))///
| (effnsum < 68 & !missing(nsum))*/
```

Statistical Tests of Differences

As described in Section 7 of the report, significance tests were conducted on differences of prevalence estimates between the 2013 NSDUH and previous years of NSDUH back to 2002, as well as differences of prevalence estimates between combined 2010–2011 survey data and combined 2012–2013 survey data. Note that for year-to-year tests of differences, if the estimate for either year is suppressed, then the resulting *p* value is also suppressed. This is the rule used when creating the detailed tables and mental health detailed tables; however, this code does not show this rule being implemented.

For the SUDAAN example (Exhibit A.7), 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.7) uses the DIFFVAR statement to test for differences between the 2012 and 2013 past month alcohol use estimates for all people aged 12 or older (IRSEX=0), all males (IRSEX=1), and all females (IRSEX=2). Similarly, for the Stata example (Exhibit A.8), a separate svy: mean command is needed.

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 controlled domain or an uncontrolled domain. Both ways are described as follows with accompanying example code: Exhibits A.7 and A.9 show example code for uncontrolled domains using SUDAAN and SAS, and Exhibits A.8 and A.10 show the same examples using Stata. Exhibits A.7, A.11, A.13, A.15, and A.17 show example code for controlled domains using SUDAAN and SAS and Exhibits A.8, A.12, A.14, A.16, and A.18 show the same examples using Stata.

Exhibit A.7 SUDAAN DESCRIPT Procedure (Tests of Differences)

```
PROC DESCRIPT DATA=DATANAME DDF=900 DESIGN=WR FILETYPE=SAS;
NEST VESTR VEREP;
WEIGHT ANALWT;
VAR ALCMON;
SUBGROUP YEAR IRSEX;
LEVELS 2 2;
TABLES IRSEX;
DIFFVAR YEAR=(1 2); /*Tests of differences between 2011(year=1)
                      and 2012 (year=2)*/
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”;
TITLE "TESTS OF DIFFERENCES BETWEEN 2012 AND 2013 ESTIMATES OF
PAST MONTH ALCOHOL BY YEAR AND GENDER";
RUN;
```

```
Note: The following CLASS statement could be used in place of SUBGROUP
and LEVELS statements in the above example:
      CLASS YEAR IRSEX;
```

When one or more contrasts are specified in SUDAAN, as in the DIFFVAR statement above, the output variable MEAN becomes the contrast mean, and SEMEAN becomes the SE of the contrast mean. The example above also outputs the *t*-statistic (T_MEAN) and the corresponding *p* value (P_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). With these statistics and the correct degrees of freedom (900 in this example), the *p* value (PVALT) for the test of differences between totals for uncontrolled domains can be calculated as indicated in Exhibit A.9. The SAS function PROBT returns the probability from a *t*-distribution.

Exhibit A.8 Stata COMMANDS svy: mean and svy: total (Tests of Differences)

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

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

/*ID Nesting variables (VESTR and VEREP) and weight variable
(ANALWT - standard single-year, person-level analysis weight*/
svyset verep [pweight=analwt], strata(vestr) dof(900)
{
svy: mean alcmon, over(year irsex)
```

Exhibit A.8 Stata COMMANDS svy: mean and svy: total (Tests of Differences) (continued)

```
local max=2*2 /*number of years*number of gender categories. This
is the total number of supops*/
local range=2 /*number of gender categories. This is the number
of subpops per year*/
local compmin=`max'-`range'
gen pmean=. /*P-value T-test Cont. Mean=0*/
local counter=1
forvalues i=1/1 { /*number of contrasts needed to compare year==1
vs year==2*/
      local counter2=1
      forvalues j=1/2 { /*number of gender categories*/
            local stop=`counter2'+`compmin'
            test [alcmon]_subpop_`counter' = ///
            [alcmon]_subpop_`stop', nosvyadjust
            replace pmean=r(p) if year==`i' & irsex==`j' /*p-value
      t-test cont. mean=0*/
            local counter=`counter'+1
            local counter2=`counter2'+1
      }
            }
      }

svy: total alcmon, over(year irsex)
      {
matrix M = e(b) /*The totals for each subpopulation are stored in
here*/
local max=2*2     /*number of years*number of gender categories.
This is the total number of supops*/
local range=2    /*number of gender categories. This is the number
of subpops per year*/
local compmin=`max'-`range'
gen total_out=. /*Contrast total*/
gen setotal=.    /*Total Standard error*/
      local counter=1
      forvalues i=1/1 {  /*number of contrasts needed to compare
year==1 vs year==2*/
            local counter2=1
            forvalues j=1/2 {     /*number of gender categories*/
                  local stop=`counter2'+`compmin'
                  test [alcmon]_subpop_`counter' = ///
      [alcmon]_subpop_`stop', nosvyadjust matvlc(test`counter')

                  replace setotal= sqrt((test`counter'[1,1])) ///
      if year==`i' & irsex==`j'
```

Exhibit A.8 Stata COMMANDS svy: mean and svy: total (Tests of Differences) (continued)

```
                replace total=M[1,`counter']-M[1,`stop'] ///
      if year==`i' & irsex==`j' /*Calculating the difference
between the totals of the subpopulation*/
                local counter=`counter'+1
                local counter2=`counter2'+1
                }
           }
      }

      *Keeping variables that matches SUDAAN
      keep irsex total setotal pmean
      duplicates drop irsex total setotal pmean, force /*keep one
record per contrast*/

      drop if total_out == . /* drop the rows where there is no
information */
      format pmean %-15.10f
      format total_out %-12.0fc
      format setotal %-12.0fc

      /* Output the dataset*/
      list irsex total_out setotal pmean
```

Exhibit A.9 SAS Code (Calculation of the *P* Value for the Test of Differences between Totals for Uncontrolled Domains)

```
IF SETOTAL GT 0.0 THEN DO; /*SETOTAL and TOTAL come from Exhibit
A.7*/
   PVALT=2*(1-PROBT(ABS(TOTAL/SETOTAL),900));
END;
```

Exhibit A.10 Stata Code (Calculation of the *P* Value for the Test of Differences between Totals for Uncontrolled Domains)

```
generate pvalt = tprob(900,abs(total_out /setotal)) ///
if setotal > 0 & !missing(setotal) /* two-tail*/
/*total_out and setotal come from Exhibit A.8*/
```

In Exhibits A.1 and A.2, all people aged 12 or older and gender are annually controlled totals. For controlled domains like these, additional steps are needed to compute similar *p* values for tests of differences. One approach uses an additional DESCRIPT procedure in SUDAAN to output the appropriate covariance matrix (Exhibit A.11), and an additional svy: mean command in Stata outputs a similar matrix (Exhibit A.12). Then, through further SAS or Stata data manipulations, the weighted sample sizes (WSUM), variances, and the covariance of the two means (obtained from the covariance matrix) are used to generate the standard *t* test statistic. The corresponding *p* value can once again be produced using the SAS PROBT function or Stata TPROB function and calculated *t* test statistic.

Exhibit A.11 SUDAAN DESCRIPT Procedure (Covariance Matrix)

```
PROC DESCRIPT DATA=DATANAME DDF=900 DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR VEREP;
WEIGHT ANALWT;
VAR ALCMON;
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;
```

```
Note: The following CLASS statement could be used in place of SUBGROUP
and LEVELS statements in the above example:
      CLASS YEAR IRSEX;
```

Exhibit A.12 Stata COMMAND svy: mean (Covariance Matrix)

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

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

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

svyset verep [pweight=analwt], strata(vestr) dof(900)
svy: mean alcmon, over(year irsex)
*Save and display the Covariance Matrix
matrix M = e(V)
matrix list M
```

The covariances of the estimated means can be obtained from the output of the DESCRIPT procedure (Exhibit A.11) and svy: mean command (Exhibit A.12). The covariance matrix in SUDAAN consists of a row and column for each gender (total, male, female) and year (both years, 2012, and 2013) combination with each cell corresponding to a particular variance component (i.e., a 9 x 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. The Stata matrix would look similar but with a few exceptions: total rows and columns would not be included (i.e., year=0 and irsex=0) and the order would be reversed (i.e., year would be listed first, followed by irsex). Table A.3 presents the Stata matrix shell.

Table A.2 SUDAAN Matrix Shell

			IRSEX=0			IRSEX=1			IRSEX=2		
			YEAR=0	YEAR=1	YEAR=2	YEAR=0	YEAR=1	YEAR=2	YEAR=0	YEAR=1	YEAR=2
		ROWNUM	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									

Table A.3 Stata Matrix Shell

OVER:	YEAR	IRSEX
_subpop_1:	1	1
_subpop_2:	1	2
_subpop_3:	2	1
_subpop_4:	2	2

	alcmon: _subpop_1	alcmon: _subpop_2	alcmon: _subpop_3	alcmon: _subpop_4
alcmon:_subpop_1				
alcmon:_subpop_2				
alcmon:_subpop_3				
alcmon:_subpop_4				

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 either 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 in Exhibit A.13.

The three SAS datasets created by the following examples, one containing the covariances (Exhibit A.13) and two containing the variances (Exhibit A.15), are then merged with the output dataset from the DESCRIPT procedure that generated the tests of differences (Exhibit A.7). With the proper statistics contained in one dataset, the corresponding *p* value for the tests of differences between controlled totals can be produced using the SAS PROBT function and calculated *t* test statistic (Exhibit A.17). Interwoven in with these three SAS code examples are Exhibits A.14, A.16, and A.18, which show Stata code performing the same functions.

Exhibit A.13 SAS Code (Identification of Covariance Components)

```
DATA COV(KEEP=IRSEX COV1);
  SET OUT.SUDCOV;
    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;
```

Exhibit A.14 Stata Code (Identification of Covariance Components)

```
local max=2*2         /*number of years*number of gender
categories. This is the total number of supops*/
local range=2         /*number of gender categories. This is the
number of subpops per year*/
local compmin=`max'-`range'

gen cov1=1
local counter=1
forvalues i=1/1 {  /*number of contrasts needed to compare year=1
vs year=2*/
     local counter2=1
     forvalues j=1/2 {     /*number of gender categories*/
          local stop=`counter2'+`compmin'
          replace cov1=M[`j', `stop'] if irsex==`j'
          local counter=`counter'+1
          local counter2=`counter2'+1
          }
     }

duplicates drop irsex cov1, force
     list irsex cov1
     keep irsex cov1
/* Save data to network*/
save ".\\cov.dta" , replace  /*Need to save dataset since Stata
can only work with one at a time*/
```

The variances of the means are calculated in separate data steps shown in Exhibits A.15 and A.16. The variance is simply the square of the SE of the mean. The SEs of the means were output in the original procedure that generated the estimates (DESCRIPT for the SUDAAN/SAS example and svy: mean for the Stata example; see Exhibits A.1 and A.2).

Exhibit A.15 SAS Code (Calculation of Variances)

```
    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;
```

Exhibit A.16 Stata Code (Calculation of Variances)

```
/*Run code from Exhibit A.2 or save the output from that exhibit
into a dataset then read in that dataset here then run the
remaining code.*/
/*Note: The remaining code for this exhibit will need to be run as
a block to avoid errors.*/
preserve  /*keep dataset in memory*/


keep if year ==1
gen wsum1 = wsum
gen var1  = semean^2
keep wsum1 var1 year irsex

duplicates drop year irsex, force /*keep one record per
subpopulation of interest*/

save ".\\est1.dta" , replace  //Need to save dataset since Stata
could only work with one at a time

restore, preserve /*restore dataset back to normal and edit for
second dataset*/


keep if year==2
gen wsum2 = wsum
gen var2  = semean^2
keep wsum2 var2 year irsex

duplicates drop year irsex, force /*keep one record per
subpopulation of interest*/

save ".\\est2.dta" , replace  /*Need to save dataset since Stata
could only work with one dataset at a time*/

restore, preserve
```

Exhibit A.17 SAS Code (Calculation of the *P* Value for the Test of Differences between Totals for Controlled Domains)

```
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)),900));
RUN;
```

Exhibit A.18 Stata Code (Calculation of the *P* Value for the Test of Differences between Totals for Controlled Domains)

```
/*Run code from Exhibit A.8, A.14, and A.16 then run the
remaining code to calculate the p values*/

keep irsex total_out

*merge by irsex for dataset est1 est2 cov
merge m:m irsex using ".\\est1.dta", generate(_merge1)
merge m:m irsex using ".\\est2.dta", generate(_merge2)
merge m:m irsex using ".\\cov.dta", generate(_merge3)
generate pvalt = tprob(900,abs(total_out ///
/sqrt(wsum1^2*var1+wsum2^2*var2-2*wsum1*wsum2*cov1))) /*
 two-tail*/

drop _merge1 _merge2 _merge3
list irsex year wsum1 var1 wsum2 var2 cov1 pvalt
```

Recoding and Missing Values

In the example in Exhibit A.19 (using SAS and SUDAAN) and A.20 (using Stata), the mean age of first use of marijuana will be calculated in two ways within 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 the calculation of mean age of first use of marijuana using SAS and SUDAAN or Stata.

Exhibit A.19 SAS Code (Recoding a Variable) and SUDAAN DESCRIPT Procedure (Estimate Generation with (1) Missing Values and (2) Using Subpopulation)

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

```
DATA DATANAME;
 SET DATANAME;
 IF IRMJAGE=991 THEN IRMJAGE_R=.;
 ELSE IRMJAGE_R=IRMJAGE;
 RUN;
```

Exhibit A.19 SAS Code (Recoding a Variable) and SUDAAN DESCRIPT Procedure (Estimate Generation with (1) Missing Values and (2) Using Subpopulation) (continued)

```
 PROC DESCRIPT DATA=DATANAME DDF=900 DESIGN=WR
 FILETYPE=SAS DEFT4;
 NEST VESTR VEREP;
 WEIGHT ANALWT;   /*Standard single-year, person-level
 analysis weight*/
 VAR IRMJAGE_R;   /*Marijuana Age of First Use recoded
 analysis variable*/
 SUBGROUP IRSEX;
 /*Gender variable, where male=1 & female=2*/
 LEVELS 2;
 TABLES IRSEX; /*Gender*/
 PRINT MEAN SEMEAN / REPLACE STYLE=NCHS;
 TITLE "ESTIMATES OF AGE OF FIRST USE OF MARIJUANA BY
 GENDER";
RUN;
```

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

```
PROC DESCRIPT DATA=DATANAME DDF=900 DESIGN=WR FILETYPE=SAS DEFT4;
NEST VESTR VEREP;
WEIGHT ANALWT;   /*Standard single-year, person-level analysis
weight*/
SUBPOPN MRJFLAG=1; /*Sub setting 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;
     /*Gender variable, where male=1 & female=2*/
LEVELS 2;
TABLES IRSEX; /*Gender*/
PRINT MEAN SEMEAN / REPLACE STYLE=NCHS;
TITLE "ESTIMATES OF AGE OF FIRST USE OF MARIJUANA BY GENDER";
RUN;
```

Exhibit A.20 Stata Code (Recoding a Variable, Estimate Generation with (1) Missing Values and (2) Using Subpopulation)

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

generate irmjage_r = irmjage
replace irmjage_r = . if irmjage == 991
```

Exhibit A.20 Stata Code (Recoding a Variable, Estimate Generation with (1) Missing Values and (2) Using Subpopulation) (continued)

```
/*Method 1, recoding unused values to missing*/
svyset verep [pweight=analwt], strata(vestr) dof(900)
svy: mean  irmjage_r, over(irsex)
/*marijuana age of first use analysis variable, gender variable*/

/*Method 2, using subpopulation to omit the unused values*/
svyset verep [pweight=analwt], strata(vestr) dof(900)
svy, subpop(mrjflag): mean  irmjage, over(irsex)
```

Confidence Intervals

As discussed in Section 8 of this report, confidence intervals can be calculated using means (MEAN) and SEs (SEMEAN) from PROC DESCRIPT in SUDAAN or svy: mean in Stata. After the means and standard errors are obtained (Exhibits A.1 and A.2), the code in Exhibits A.21 and A.22 can be used to create the 95 percent confidence intervals.

Exhibit A.21 SAS Code (Calculating a 95 Percent Confidence Interval for a Mean)

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

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

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

          RUN;
```

Exhibit A.22 Stata Code (Calculating a 95 Percent Confidence Interval for a Mean)

```
/*Run code from Exhibit A.2 or save output dataset from Exhibit
A.2 and use that as input to this code.*/
generate t_qntile = invt(900,0.975)
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))
```

Exhibit A.22 Stata Code (Calculating a 95 Percent Confidence Interval for a Mean) (continued)

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

duplicates drop year irsex, force /*keep one record per
subpopulation of interest*/
keep year irsex nsum wsum mean_out semean  t_qntile number ///
l a b plower pupper
```

Calculating Percentages for Categories

Exhibits A.23 and A.24 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 (MRJDAYS) 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. Note that the suppression rule for percentages is the same as the suppression rule for means shown in Exhibit A.5, except PERCENT and SEPERCENT have to be divided by 100 (and thus are equivalent to MEAN and SEMEAN in the formulas). 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.

Exhibit A.23 SAS Code (Frequency of Use, i.e., Number of Days Used Substance in the Past Month among Past Month Users)

```
PROC DESCRIPT DATA=DATANAME DDF=900 DESIGN=WR FILETYPE=SAS DEFT4;
   NEST VESTR VEREP;
   WEIGHT ANALWT;   /*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*/
   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 percents among marijuana
   users*/
   PRINT WSUM NSUM PERCENT SEPERCENT TOTAL SETOTAL / REPLACE
   STYLE=NCHS;
   OUTPUT WSUM PERCENT SEPERCENT TOTAL SETOTAL NSUM  /REPLACE
   FILENAME="OUT.SUDFILE_FREQ";
   TITLE "FREQUENCY OF MARIJUANA USE BY PAST MONTH MARIJUANA
USERS";RUN;
```

Exhibit A.24 Stata Code (Frequency of Use, i.e., Number of Days Used Substance in the Past Month among Past Month Users)

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

      svyset verep [pw=analwt], strata(vestr) dof(900)
      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*/
```

Testing Between Overlapping Domains

In addition to testing between-year differences shown in Exhibits A.7 and A.8, Exhibits A.25 and A.26 demonstrate testing between two overlapping domains. 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 people aged 18 or older who are employed full time.

This code will apply when one domain is completely contained in another or when there is only partial overlap. The example below uses two domains, where one domain is completely contained in the other (i.e., comparing unemployed adults to all adults—the unemployed group is completely contained by the all adults group). Note that the correlations between the two estimates are accounted for in this test (i.e., correlation between past month cigarette use among people aged 18 or older and past month cigarette use among people aged 18 or older employed full time).

Exhibit A.25 SAS Code (Test of Difference when Two Groups Overlap Using Stacked Data)

```
      DATA STACKED;
            SET DATANAME(IN=A) DATANAME(IN=B); /*reading in data
      twice*/
            IF A THEN DO;
      INDIC=1;
      IF EMPSTAT4 IN (1,2,3,4) THEN EMPLOY=1;
      /*EMPSTAT4 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 EMPSTAT4=1 THEN EMPLOY=1;
      ELSE EMPLOY=0;
            END;
```

Exhibit A.25 SAS Code (Test of Difference when Two Groups Overlap Using Stacked Data) (continued)

```
/*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*/
RUN;

PROC SORT DATA=STACKED;
BY VESTR VEREP;
RUN;
PROC DESCRIPT DATA=STACKED DDF=900 DESIGN=WR FILETYPE=SAS;
NEST VESTR VEREP;
WEIGHT ANALWT;
VAR CIGMON;
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”;
TITLE "TESTS OF DIFFERENCES BETWEEN ALL PERSONS 18 OR OLDER AND
EMPLOYED PERSONS 18 OR OLDER";
RUN;
```

Exhibit A.26 Stata Code (Test of Difference when Two Groups Overlap Using Stacked Data)

```
/*Creating the first dataset*/
/*Read in data */
use using ".\\dataname.dta", clear
/*Ensure all variables are lower case*/
rename *, lower

gen indic = 1
gen employ = 0
replace employ = 1 if inlist(empstat4,1,2,3,4)
/*Save the dataset*/
save ".\\a26_a.dta" , replace  /*Need to save dataset since Stata
can only work with one at a time*/

/*Creating the second dataset*/
/*Read in data a second time*/
use using ".\\dataname.dta", clear
```

Exhibit A.26 Stata Code (Test of Difference when Two Groups Overlap Using Stacked Data) (continued)

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

gen indic = 2
gen employ = 0
replace employ = 1 if inlist(empstat4,1)
*Save the dataset
save ".\\a26_b.dta" , replace  /*Need to save dataset since Stata
could only work with one at a time*/

/*Need to stack the dataset together */
use using ".\\a26_a.dta", clear
append using ".\\a26_b.dta"

/*Create the subpopulation variable*/
generate subpop = 1 if catag18 == 1 & employ == 1
svyset verep [pweight=analwt], strata(vestr) dof(900)
svy, subpop(subpop): mean cigmon, over(indic)
test [cigmon]1 = [cigmon]2
/*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 0 otherwise    */
```

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). 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 SUDAAN (using code similar to Exhibit A.25) or Stata (Exhibit A.26). Exhibits A.27 and A.28 show the code for calculating the Wald *F* test to determine whether cigarette use is associated with employment status.

Exhibit A.27 SAS Code (Test for Independence Based on a Log-Linear Model)

```
PROC CROSSTAB DATA=DATANAME DDF=900 DESIGN=WR FILETYPE=SAS DEFT4;
   NEST VESTR VEREP;
   WEIGHT ANALWT;
   CLASS CIGMON;
   SUBGROUP EMPSTAT4; /*four level employment status variable*/
   LEVELS   4;
   SETENV DECWIDTH=6 COLWIDTH=17;
   TABLES   EMPSTAT4*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.28 Stata Code (Test for Independence Based on a Log-Linear Model)

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


/*Need to subset to just 4 levels of empstat4*/
generate subpop = 1 if inlist(empstat4,1,2,3,4)
/*four level employment status variable*/

svyset verep [pw=analwt], strata(vestr) dof(900)

svy, subpop(subpop): tab cigmon empstat4, llwald noadjust

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