2016-2018 National Survey on Drug Use and Health: Comparison of Population Percentages from the United States, Census Regions, States, the District of Columbia, and Substate Regions (Documentation for CSV and Excel Files)

This page intentionally left blank

Documentation for CSV and Excel Files

Description of the CSV File Type

Files with a comma separated value (*.csv) extension are in plain text. They contain characters stored in a flat, nonproprietary format and can be opened by most computer programs. Each *.csv file contains a set of tabular data, with each record delineated by a line break and each field within a record delineated by a comma. A field that contains commas as part of its content has the additional delineation of a quote mark character before and after the field's contents. When a quote mark character is part of a field's content, it is included as two consecutive ""quote mark"" characters.

Computers with Microsoft Excel installed open *.csv files in Excel by default, with the fields automatically arranged appropriately in columns. Other database programs also open *.csv files with the fields appropriately arranged.

The 1,664 CSV files (i.e., "NSDUHsubstatePValueTab#-#_2018.csv") reflect the 1,664 Excel tables,[1] and they contain the table title, table notes, column headings, and data. The webpage at https://www.samhsa.gov/data/ for the 2016-2018 NSDUH state *p* value tables includes a hyperlinked table of contents on the first sheet of the Excel file that combines all of the Excel tables, as well as a ZIP file containing all of the individual CSV files. Additionally, the ZIP file includes a text file with a list of the table numbers and titles.

How to Use the *P* Value Tables

The *p* values contained in the national tables (the state tables are a subset of the national tables) for each outcome can be used to test the null hypothesis of no difference between population percentages for the following types of comparisons:

- *total United States versus census region* (within the table, to find the *p* value, go to the census region row, then to the Total United States column),
- *total United States versus state* (within the table, to find the *p* value, go to the state row, then to the Total United States column),
- *total United States versus substate region* (within the table, to find the *p* value, go to the substate region row, then to the Total United States column),
- *census region versus census region* (within the table, to find the *p* value, go to the second census region's row, then navigate to the column of the other census region),
- *census region versus state* (within the table, to find the *p* value, go to the state row, then navigate to the census region column),

[1] These include 32 tables for each of the 50 states and the District of Columbia, presenting *p* values for the substate regions in each state (see Tables 1.1 to 51.32), and 32 national tables that include estimates for all of the substate regions in the United States (see Tables 1 to 32).

- *census region versus substate region* (within the table, to find the *p* value, go to the substate region row, then navigate to the census region column),
- *state versus state* (within the table, to find the *p* value, go to the second state's row, then navigate to the column of the other state),
- *state versus substate region* (within the table, to find the *p* value, go to the second state's row or the substate region's row, then navigate to the column of the other area), and
- *substate region versus substate region* (within the table, to find the *p* value, go to the second substate region's row, then navigate to the column of the other substate region).

For example, within any given national table (in Tables 1 to 32), by scrolling across Alabama's state *row* to the South's census region *column*, the *p* value found will determine whether Alabama's state population percentage and the South's census region population percentage are significantly different for a particular outcome of interest. By scrolling across the row for Alabama's Region 2 to the column for the state of Alabama, the *p* value found will determine whether Alabama's state population percentage and Alabama's Region 2 population percentage are significantly different for a particular outcome of interest (the same can also be found in any Alabama table). Similarly, by scrolling across the row for Colorado's Region 7 to the column for Region 1 of Colorado in any national table, the *p* value found will determine whether Region 7's population percentage and Region 1's population percentage are significantly different for a particular outcome of interest (the same can also be found in any Colorado table). Note that the tests included here are for a given outcome and age group.[2]

The following example describes how to test the null hypothesis of no difference between population percentages in the *national file*. Table 13 in the national file contains *p* values for past month alcohol use among individuals aged 12 or older for all of the substate regions in the United States and for each state and census region. To find the *p* value for testing the null hypothesis of no difference between population percentages for past month alcohol use between two large metropolitan areas, such as Region 11 (Los Angeles) in California and Region 2: New York City in New York, scroll to the row for Region 2: New York City and navigate to the column for Region 11 (Los Angeles). That *p* value is 0.741. Thus, the hypothesis of no difference—that is, the Region 11 (Los Angeles), California, population percentage for past month alcohol use is the same as the Region 2: New York City, New York, population percentage for past month alcohol use—is not rejected at the 5 percent level of significance, meaning that the two prevalence rates are not statistically different. Note that the small area estimates for Region 11 (Los Angeles), California, and Region 2: New York City, New York, are 49.15 and 49.59 percent, respectively.[3] Similarly, to find the *p* value for testing the null hypothesis of no difference between population percentages for past month alcohol use between

[2] The types of outcomes in these tables include illicit drug use, alcohol use, tobacco use, perception of great risk of substance use, substance use disorder, needing but not receiving treatment, serious mental illness, any mental illness, mental health services, suicidal thoughts and behavior, and major depressive episode. The age groups include individuals aged 12 or older, adults aged 18 or older for the mental health outcomes, and individuals aged 12 to 20 for underage alcohol use.

[3] See Table 13 of the "2016-2018 NSDUH Substate Region Estimates: Excel Tables and CSV Files" at https://www.samhsa.gov/data/.

two adjacent substate regions that are in two different states (e.g., the Badlands region in North Dakota and Region 1 in South Dakota), they can use the national file to find the *p* value (i.e., 0.503).

The following example describes how to test the null hypothesis of no difference between population percentages in a *state file*. Table 2.3 in Alaska's file contains *p* values for past month marijuana use among individuals aged 12 or older for the four substate regions in Alaska and for the West census region. The *p* value for testing the null hypothesis of no difference between Anchorage, Alaska, and the West census region population percentages for past month marijuana use is 0.003, which is found by scrolling to the row for Anchorage and navigating to the column for the West census region. Thus, the hypothesis of no difference (i.e., the Anchorage population percentage for past month marijuana use is the same as the West census region population percentage for past month marijuana use) is rejected at the 5 percent level of significance, meaning that the two prevalence rates are statistically different. Note that the Anchorage, Alaska, and the West census region estimates are 15.62 and 12.39 percent, respectively.[4]

Comparison between Two Small Area Population Percentages

To produce substate region, state, census region, and national small area estimates, the 2016-2018 NSDUH data were modeled using the method discussed in Section B.1 of the "2016-2018 NSDUH: Guide to Substate Tables and Summary of Small Area Estimation Methodology" document at https://www.samhsa.gov/data/. This modeling results in 1,250 Markov Chain Monte Carlo (MCMC) samples that are used here to calculate *p* values for testing the null hypothesis of no difference between two small area population percentages.

To test the difference between the population percentages of two areas, let π_{1a} and π_{2a} denote the 2016-2018 population percentages of two areas (e.g., state 1 vs. state 2 or state 1 vs. national, or substate region 1 vs. substate region 2) for age group-*a*. The null hypothesis of no difference, $H_0 : \pi_{1a} = \pi_{2a}$, is equivalent to the null hypothesis, $H_0 : lor_a = 0$, where lor_a is the log-odds ratio, $lor_a = \ln\left[\frac{\pi_{2a}/(1-\pi_{2a)})}{\pi_{1a}/(1-\pi_{1a})}\right]$, and ln denotes the natural logarithm. The difference between π_{1a} and π_{2a} is defined in terms of the log-odds ratio as opposed to the simple difference $(\pi_{2a} - \pi_{1a})$ because the posterior distribution of the log-odds ratio is closer to Gaussian than the posterior distribution of the simple difference.

An estimate, $\hat{lor}_a$, of lor_a is given by the average of the 1,250 MCMC sample-based log-odds ratios. Let $lor_a(i)$ denote the log-odds ratio for the *i*-th MCMC sample. That is,

$$lor_a(i) = \ln\left[\frac{\pi_{2a}(i)/[1-\pi_{2a}(i)]}{\pi_{1a}(i)/[1-\pi_{1a}(i)]}\right], i = 1, \cdots, 1,250.$$

[4] See Table 3 of the "2016-2018 NSDUH Substate Region Estimates: Excel Tables and CSV Files" at https://www.samhsa.gov/data/.

Then $\hat{lor}_a = [\sum_{i=1}^{1,250} lor_a(i)]/1,250,$ and the posterior variance of lor_a is given by $v(\hat{lor}_a) = [\sum_{i=1}^{1,250} (lor_a(i) - \hat{lor}_a)^2]/1,250.$

For testing the null hypothesis of no difference $(lor_a = 0)$, it is assumed that the posterior distribution of lor_a is normal with $mean = \hat{lor}_a$ and $variance = v(\hat{lor}_a)$. The Bayesian *p* value or significance level for the null hypothesis of no difference $(lor_a = 0)$ is p value $= 2 * P[Z \geq abs(z)]$, where Z is a standard normal random variate, $z = \hat{lor}_a / sqrt[v(\hat{lor}_a)]$, and $abs(z)$ denotes the absolute value of z. This Bayesian significance level (or *p* value) for the null value of lor, say lor_0, is defined following Rubin[5] as the posterior probability for the collection of the lor values that are less likely or have smaller posterior density $d(lor)$ than the null (no change) value lor_0. That is, p value$(lor_0) =$ probability$[d(lor_a) \leq d(lor_0)]$. With the posterior distribution of lor approximately normal, p value(lor_0) is given by the above expression. If the *p* value is less than 0.05, then it can be stated that the estimates for the two areas are statistically different from each other.

[5] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics). New York, NY: John Wiley & Sons.