2016-2017 National Survey on Drug Use and Health: Comparison of Population Percentages from the United States, Census Regions, States, and the District of Columbia (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 143 CSV files (i.e., "P Value Table#.csv") reflect the 143 Excel tables, and they contain the table title, table notes, column headings, and data. The webpage at https://www.samhsa.gov/data/ for the 2016-2017 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 these tables for each outcome and age group 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 U.S. column),
- *total United States versus state* (within the table, to find the *p* value, go to the state row, then to the Total U.S. column),
- *census region versus census region* (within the table, to find the *p* value, go to the census region with the higher order number, then navigate to the column of the other 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), and
- *state versus state* (within the table, to find the *p* value, go to the state row with the higher order number [i.e., the state that is lower alphabetically], then navigate to the column of the other state [i.e., the state that is higher alphabetically]).

In general, to find the *p* value when testing any two geographic areas, navigate to the row of the area with the higher order number, then navigate to the column of the other area. For example,

within any given table, 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. Note that the tests included here are for a given outcome and age group.[1]

For example, Table 2.2 contains *p* values for past year marijuana use among youths aged 12 to 17. The *p* value for testing the null hypothesis of no difference between Oregon and the West region population percentages for past year marijuana use among youths aged 12 to 17 is 0.007. Thus, the hypothesis of no difference (Oregon population percentage = West region population percentage) is rejected at the 5 percent level of significance, meaning that the two prevalence rates are statistically different. Note that the Oregon and West region estimates for youths aged 12 to 17 are 17.0 and 13.6 percent, respectively.[2]

Comparison between Two Small Area Population Percentages

To produce state, census region, and national small area estimates, the 2016-2017 NSDUH data were modeled using the method discussed in Section B.1 of the "2016-2017 NSDUH: Guide to State 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.

Let π_{1a} and π_{2a} denote the 2016-2017 population percentages of two areas (e.g., state 1 vs. state 2 or state 1 vs. national) for age group-a. The difference between π_{1a} and π_{2a} is defined in terms of the log-odds ratio, $lor_a = \ln\left[\dfrac{\pi_{2a}/(1-\pi_{2a})}{\pi_{1a}/(1-\pi_{1a})}\right]$, where ln denotes the natural logarithm, 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,

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

[1] The outcomes in these tables focus on 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, youths aged 12 to 17, young adults aged 18 to 25, adults aged 26 or older, and adults aged 18 or older. Alcohol use is also provided for individuals aged 12 to 20 (i.e., underage individuals). Note that not all outcomes have data broken out by all age groups.

[2] See Table 2 of the "2016-2017 NSDUH: Model-Based Prevalence Estimates (50 States and the District of Columbia)" at https://www.samhsa.gov/data/.

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

$$v(\hat{lor}_a) = [\sum_{i=1}^{1,250} (lor_a(i) - \hat{lor}_a)^2] / 1,250.$$

To calculate the *p* value 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 $variance = v(\hat{lor}_a).$ With $(lor_a = 0),$ the Bayes *p* value or significance level for the null hypothesis of no difference 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[3] 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, $p\,value(lor_0) = probability[d(lor) \leq d(lor_0)]$. 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 population percentages of two areas are statistically different from each other.

[3] 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.

This page intentionally left blank