

**2023-2024 National Surveys on Drug Use
and Health: Guide to Comparison of
United States, Census Regions, States,
and the District of Columbia Population
Percentages (Documentation for CSV
and Excel Files)**

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 181 *.csv files (i.e., "P Value Table#.csv") reflect the 181 Excel tables, and they contain the table title, table notes, column headings, and data. The 2023-2024 National Surveys on Drug Use and Health (NSDUHs) state *p* value tables include 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. These files are available on the [NSDUH Data Collection](#) web page. 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 ["Order" is the first column in the Excel tables and includes values 1 (Total U.S.) to 56 (Wyoming)]),
- *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, then navigate to the column of the other state ["Order" is the first column in the Excel tables and includes values 1 (Total U.S.) to 56 (Wyoming)]).

In general, to find the p value for testing the difference between the population percentage of 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, for a given outcome and age group, are produced using 2023 and 2024 data.¹

For example, Table 2.3 contains the p values for testing the difference between the population percentage of any two geographic areas for marijuana use in the past year among people aged 18 to 25. It can be seen, for example, that the p value for testing the null hypothesis of no difference between the South region and Alabama population percentages for marijuana use in the past year for people aged 18 to 25 is 0.002 (Table 2.3, row = Alabama, column = South). Thus, the hypothesis of no difference (Alabama population percentage = South region population percentage) is rejected at the 1 percent level of significance, meaning that the two population percentages are statistically different. Note that the Alabama and South region estimates for marijuana use in the past year among people aged 18 to 25 are 24.50 and 30.87 percent, respectively.²

Comparison between Two Small Area Population Percentages

To produce state, census region, and national small area estimates, the 2023-2024 NSDUH data³ were modeled using the method discussed in Section B.1 of *2023-2024 National Surveys on Drug Use and Health: Guide to State Tables and Summary of Small Area Estimation Methodology* on the [NSDUH Data Collection](#) web page. 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 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) as opposed to the simple difference ($\pi_{2a} - \pi_{1a}$) because the posterior

¹ The outcomes in these tables focus on illicit drug use, alcohol use, tobacco use, perception of great risk of harm from substance use, substance use disorder (SUD), substance use treatment, any mental illness, serious mental illness, co-occurring SUD and mental illness, mental health treatment, major depressive episode, and suicidal thoughts and behavior. The age groups include people 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- and tobacco-related outcomes are also provided for people aged 12 to 20 (i.e., underage people). Note that not all outcomes have data broken out by all age groups. Estimates for youths aged 12 to 17 are not available for past year heroin use because past year heroin use was extremely rare among youths aged 12 to 17 in the 2023 and 2024 NSDUHs. As a result, estimates for people aged 12 or older are also not produced. Thus, p value tables for these two age groups for past year heroin use are not available.

² See Table 2 of *2023-2024 National Surveys on Drug Use and Health: Model-Based Prevalence Estimates (50 States and the District of Columbia)* on the [NSDUH Data Collection](#) web page.

³ The substance use treatment, mental health treatment, and illicit drug use other than marijuana estimates are based on data from only the 2024 NSDUH because there were no comparable data in 2023 for those measures.

distribution of lor_a is closer to Gaussian than the posterior distribution of the simple difference. Let \ln denote the natural logarithm, then lor_a is defined as follows:

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

Let $lor_a(i)$ denote the log-odds ratio for the i -th MCMC sample. That is,

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

Then an estimate of lor_a is given by the average of the 1,250 MCMC sample-based log-odds ratios, namely $\hat{lor}_a = \left[\sum_{i=1}^{1,250} lor_a(i) \right] / 1,250$, and the variance of \hat{lor}_a is given by the following:

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

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 estimated $mean = \hat{lor}_a$ and $variance = v(\hat{lor}_a)$. The Bayesian p value or significance level for the null hypothesis of no difference, ($lor_a = 0$), is $p \text{ value} = 2 * P[Z \geq abs(z)]$, where Z is a standard normal random variate, $z = \hat{lor}_a / \text{sqrt}[v(\hat{lor}_a)]$, and $abs(z)$ denotes the absolute value of z . This Bayesian significance level (or p value) for the null value of lor_a , say lor_0 , is defined following Rubin (1987)⁴ as the posterior probability for the collection of the lor_a values that are less likely or have smaller posterior density, $d(lor_a)$, than the null (no change) value, lor_0 . That is,

$$p \text{ value}(lor_0) = \text{probability}[d(lor_a) \leq d(lor_0)].$$

With the posterior distribution of lor_a approximately normal, $p \text{ value}(lor_0)$ is given by the above expression. If the p value is less than 0.01, for example, then it can be stated that the population percentages of two areas are statistically different from each other at the 1 percent level of significance.

⁴ See the following reference: Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics). John Wiley & Sons.