Predicting the Number of Respondents from each State in the 2022 American Community Survey

Richard Guo

2024-11-21

Method of Obtaining the Data

To retrieve data from the IPUMS USA database(Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rogers and Megan Schouweiler 2024), we first log in to the IPUMS website and selected "IPUMS USA." From there, we clicked "Get Data" and chose the "2022 ACS" sample under the "SELECT SAMPLE" section. To obtain state-level data, we selected "HOUSEHOLD" > "GEOGRAPHIC" and added the "STATEICP" variable to the data cart. For individual-level data, we went to the "PERSON" > "EDUCATION" section and included the "EDUC" variable. After reviewing the cart, we clicked "CREATE DATA EXTRACT" and set the "DATA FORMAT" to "csv." Finally, we submitted the extract and then downloaded and saved the file locally (e.g., "usa_00001.csv") for use in RStudio(R Core Team 2023).

Next, we count the respondents with doctoral degrees grouped by states.

Table 1: doctoral respondents

STATEICP	doctoral
1	600
2	165
3	2014
4	244
5	177
6	131

Brief Overview of the Ratio Estimators Approach

The ratio estimator is a statistical method used to estimate population totals or averages by utilizing known sample ratios. It operates by calculating the ratio of a specific characteristic (in this paper, the number of individuals holding doctoral degrees) relative to the total population in a known sample (the state of California). This ratio is then applied to other regions or groups, based on the assumption that the relationship between the characteristic and the population remains consistent across different areas. This method is particularly effective when complete population data is unavailable, but sample data provides a reliable basis for inferring proportional relationships that can be generalized.

Estimates and the Actual Number of Respondents

When assuming that California has 391,171 respondents across all educational levels, we can calculate the ratio for California. The Laplace ratio method works well when the relationship between the characteristic of interest, such as the proportion of doctoral degree holders, and the population is consistent across all units. In the example above, the ratio for California is 0.01619752. However, if this ratio is not representative of other states due to unobserved factors, the estimates derived from the method can become biased. This bias occurs when regional differences disrupt the assumption of proportionality, leading to inaccurate estimates. Factors like varying educational systems, socioeconomic conditions, or migration patterns between states can cause these distortions.

```
[1] "The California ratio that was found was: 0.0161975197547875"
```

Using the ratio estimator approach might estimate the total number of respondents for each state as follows. Assuming the ratio holds for all states, we can estimate the total respondents for every state.

The table below presents an overview of the distribution of respondents with doctoral degrees across different states in the U.S., which is crucial for the understanding of regional educational attainment trends. The first column, STATEICP, refers to the unique state or territory codes provided by the IPUMS system, where each number corresponds to a specific state. The second column, doctoral_respondents, provides the number of respondents with doctoral degrees (filtered by EDUCD == 116) in each state. For example, state code 1 shows 600 respondents with doctoral degrees, while state code 3 shows a significantly higher number with 2014 respondents. Other state codes show varying counts of respondents, such as 165 for state code 2, 244 for state code 4, and so on.

Table 2: estimated total respondents

STATEICP	doctoral	estimated_total
1	600	37043
2	165	10187
3	2014	124340
4	244	15064
5	177	10928
6	131	8088

The process of comparing it with the actual data of respondents is as follows.The table below provides a comparison between the estimated total respondents, derived using the ratio estimator method, and the actual total respondents for each state. The doctoral_respondents column represents the number of individuals with doctoral degrees in each state, while the estimated_total_respondents column shows the projected total population for each state, calculated by applying the California ratio (0.01619752) to the number of doctoral respondents in each state. The estimated values were rounded to the nearest whole number. The actual_total_respondents column reflects the true population count for each state, which was directly obtained from the data. This allows for a comparison between the estimated and actual figures. In state code 1, the actual number of respondents is 37,369, slightly higher than the estimate of 37,042. For state code 3, the actual total is 73,077, significantly lower than the estimate of 124,340. These differences illustrate the potential inaccuracies that can arise from using a ratio estimator when the relationship between the doctoral respondents and the total population is not consistent across states. This highlights the importance of accounting for regional variations in such analyses.

Table 3: comparison of the estimated and actual total respondents

STATEICP	doctoral	estimated_total	actual_total
1	600	37043	37369
2	165	10187	14523
3	2014	124340	73077
4	244	15064	14077
5	177	10928	10401
6	131	8088	6860

Explanation of Why They are Different

A simple explanation for why the estimated number of respondents and actual number of respondents differ is that the naive assumption that all states share the same doctorate-respondent to total respondent ratio is false. The ratio estimator assumes that the proportion of respondents with doctoral degrees in California is representative of the proportion in other

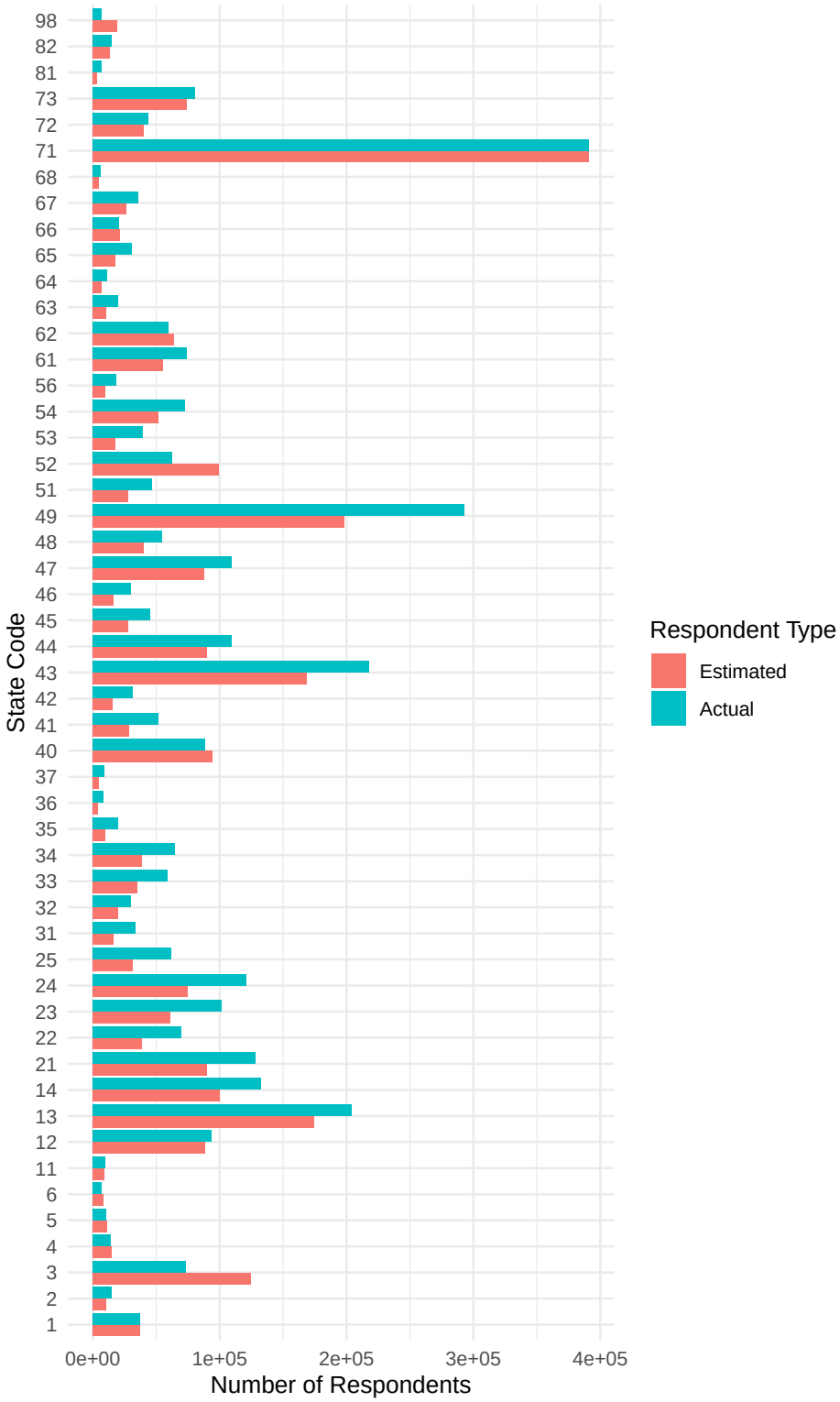

Figure 1: comparison of estimated and total respondents

states. However, educational attainment can vary significantly due to differences in demographics, economic opportunities, and educational infrastructure across states. This variance leads to discrepancies between the estimated and actual counts. When the data used for estimation is based on a sample rather than a complete population census, random sampling variability introduces uncertainty into the calculated ratio, which in turn affects the accuracy of the estimates. This variability can lead to deviations between the sample-based ratio and the true population ratio, impacting the reliability of the estimation.

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rogers and Megan Schouweiler. 2024. *IPUMS USA: Version 15.0 [dataset]*. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.