Estimating Respondents with Doctoral Degrees Using the 2022 ACS Data

Gadiel Flores, Tina Kim, and Yanfei Huang

November 21, 2024

This study aims to estimate the number of respondents with doctoral degrees across different states using the 2022 American Community Survey (ACS) dataset from IPUMS. We employ the Laplace ratio estimator, a statistical method that uses the proportion of respondents with doctoral degrees in California to estimate similar totals for other states. The analysis involves calculating the ratio of doctoral degree holders to total respondents for each state and applying this ratio to estimate state-specific totals. These estimates are then compared with the actual reported numbers. The findings underscore the importance of considering state-specific demographics and educational patterns, as discrepancies between estimated and actual totals highlight potential limitations of generalized estimation techniques. This work contributes to the broader understanding of educational attainment across the U.S. and illustrates the challenges in applying ratio estimators to diverse populations.

Table of contents

1. Introduction 2
2. Instructions for Obtaining the Data 2
3. Overview of the Ratio Estimator Approach 2
4. Data Analysis 3
5. Comparison to Actual Respondent Totals 3
6. Observations and Explanation 4
7. Conclusion 4
Appendix 5

References 5

1. Introduction

This document aims to analyze the 2022 American Community Survey (ACS) dataset from IPUMS (Center 2024). This paper uses R (R Core Team 2023) for statistical computing. To efficiently manipulate and transform the data, we utilized the dplyr package (Wickham et al. 2023). For reading in and processing large CSV files, we used the readr package (Wickham and Hester 2023). For managing file paths in our project, we used the here package (Hester 2023). We will estimate the total number of respondents in each state who have a doctoral degree using Laplace's ratio estimator (Laplace 1812). We applied a ratio estimator approach based on the methodology outlined in Cochran's book on sampling techniques (Cochran 1977). Finally, we compared these estimates to the actual number of respondents, finding that estimates may vary considerably from actual values, highlighting the need to account for state-specific demographics and contextual factors.

2. Instructions for Obtaining the Data

To access the 2022 ACS data:

1. Visit the [IPUMS website]
2. Register for an account if you haven't already.
3. Once logged in, go to the IPUMS USA section and select the 2022 ACS dataset.
4. Add the following variables to your extract:
    - `STATEICP`: State identifier based on IPUMS coding.
    - `EDUCD`: Educational attainment detail.
5. Download the `.csv.gz` file and decompress it using:

```
gunzip usa_00002.csv.gz
```

3. Overview of the Ratio Estimator Approach

The ratio estimator is a technique commonly used to estimate the size of a population based on known sample characteristics. Here, we use the following formula to estimate the total respondents in a state:

$$\text{Estimated Total Respondents} = \frac{\text{Respondents with Doctoral Degrees}}{\text{Total Respondents in the State}} \times \text{Known Total in California}$$

where:

- The numerator is the number of respondents with a doctoral degree in each state.
- The denominator is the total number of respondents in each state.
- The known total for California is provided as 391,171.

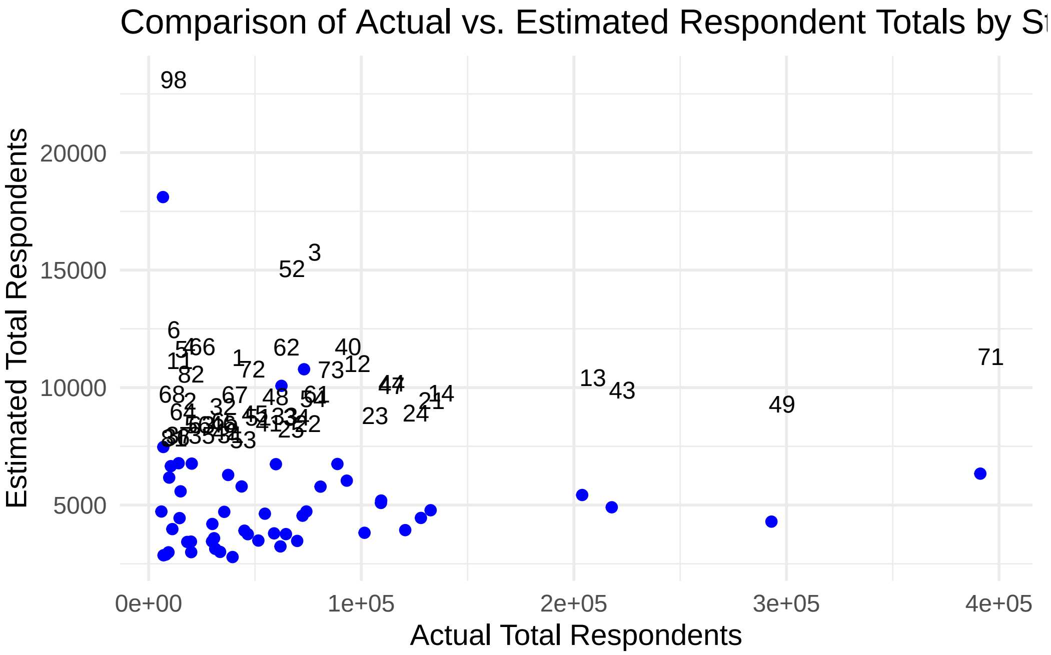
4. Data Analysis

The dataset summarizes the distribution of respondents across 51 states, focusing on those with doctoral degrees (doctoral_count) and the total number of respondents (total_count). The ratio of doctoral degree holders to total respondents varies significantly between states. Using the known total of respondents in California as a baseline, the estimated_total column extrapolates the total respondents in each state.

```
# A tibble: 51 x 5
   STATEICP doctoral_count total_count  ratio estimated_total
      <dbl>          <int>       <int>  <dbl>           <dbl>
 1        1            600       37369 0.0161           6281.
 2        2            165       14523 0.0114           4444.
 3        3           2014       73077 0.0276          10781.
 4        4            244       14077 0.0173           6780.
 5        5            177       10401 0.0170           6657.
 6        6            131        6860 0.0191           7470.
 7       11            152        9641 0.0158           6167.
 8       12           1438       93166 0.0154           6038.
 9       13           2829      203891 0.0139           5428.
10       14           1620      132605 0.0122           4779.
# i 41 more rows
```

5. Comparison to Actual Respondent Totals

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -11391   14092   41132   61048   72230  384835 
```



6. Observations and Explanation

The discrepancies between the estimated and actual total respondents in each state can be attributed to various factors:

- **Sample Representation**: The proportion of respondents with doctoral degrees may not be consistent across states, leading to differences in estimates.
- **Variation in Ratios**: States may have varying educational demographics that aren't captured accurately when using a single ratio.
- **Size of the California Sample**: The estimation depends on the accuracy of the known respondent total in California, which may not generalize well across other states.

7. Conclusion

The use of Laplace's ratio estimator provides a quick method to estimate population sizes based on sample characteristics. While useful, the results highlight that estimates can differ significantly from actual values, indicating the importance of considering state-specific demographics and context.

Appendix

The dataset and code used in this analysis are available on GitHub.

References

Center, Minnesota Population. 2024. “IPUMS USA: Integrated Public Use Microdata Series, Version 12.0 [American Community Survey 2022 Dataset].” University of Minnesota. https://usa.ipums.org/usa/.

Cochran, William G. 1977. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.

Hester, Jim. 2023. *Here: A Simple, Robust File Path Management Package*. https://CRAN.R-project.org/package=here.

Laplace, Pierre-Simon. 1812. *Théorie Analytique Des Probabilités*. Paris: Courcier.

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

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Jim Hester. 2023. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.