Estimating State-Level Doctoral Degree Attainment Using the 2022 ACS*

Yuanyi (Leo) Liu   Xuanle Zhou   Dezhen Chen   Yongqi Liu
Ziyuan Shen

October 3, 2024

This document analyzes the number of respondents in each state who have a doctoral degree as their highest educational attainment in the 2022 ACS IPUMS data. We estimate the total number of respondents in each state using the ratio estimators approach based on data from California.

Table of contents

1 Introduction

This study uses R packages (R Core Team 2020) to clean and analyze the dataset, including libraries from haven (Wickham, Miller, and Smith 2023), dplyr (Wickham et al. 2023), readr

*Code and data are available at: Estimating State-Level Doctoral Degree Attainment.

(Wickham, Hester, and Bryan 2024), kableExtra (Xie 2021), and ggplot2 (Wickham 2016). The data we used is from IPUMS (Ruggles et al. 2021).

2 A Brief Overview of Ratio Estimators Approach

The ratio estimators approach is a method used to estimate a demographic parameter such as total or mean which is using the relationship between two relevant variables. For this assignment, we selected the California population group. This group has a specific characteristic such as the proportion of the population with a doctoral degree. Then, we use the percentage of the population with a doctoral degree to estimate unknown characteristics in the general population. We can use this method when we don't know the exact size of the population, but when we can find enough sample data to assume that these proportional relationships are consistent across groups.

3 Estimates and the Actual Number of Respondents

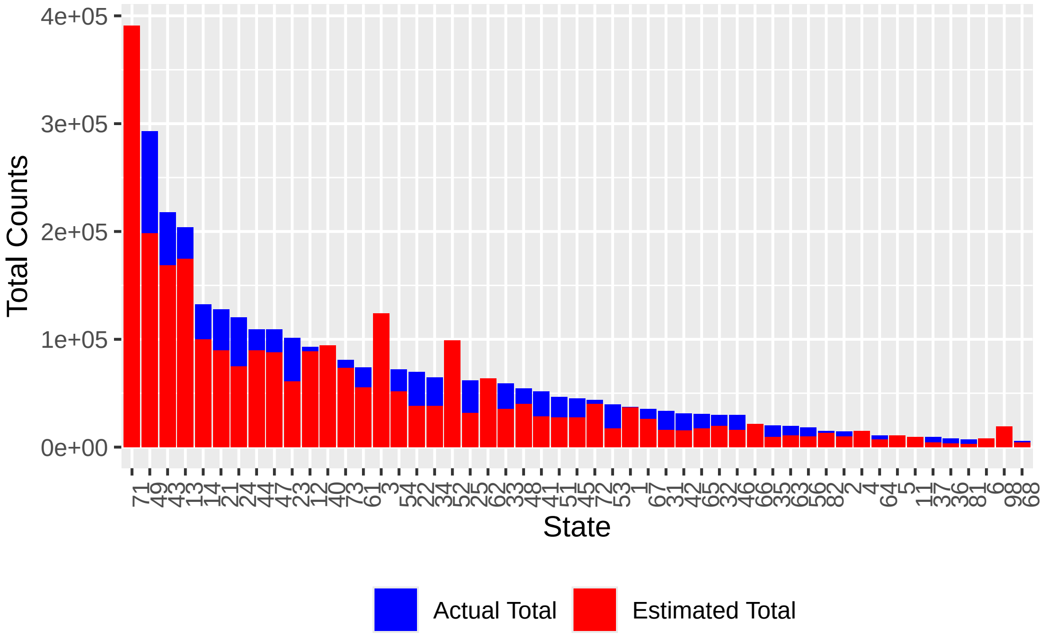

Figure 1: Comparison of Actual and Estimated Total Respondents by State

Figure 1 shows the comparison between actual and estimated respondents by state. The estimated total are higher than the actual total in most states. In particular, states in the left-hand portion of the graph show a significant difference between the actual and estimated numbers, while states in the right-hand portion of the graph show much closer numbers. This visualization allows us to clearly see the difference between estimates and actuals and helps to identify which states may need to adjust their estimation methods.

4 Explanation of the Possible Reasons for the Differences

The estimated total number of respondents in each state, calculated using the ratio estimator method, may vary from the actual count for several reasons:

Firstly, the ratio estimators approach of Laplace is based on the assumption of similarity, suggesting that the proportion of doctoral degrees in California can be representative of those in other states. However, this assumption is invalid, as numerous factors influence educational achievement. For instance, states with higher GDP per capita and stronger economic conditions typically have access to better educational resources. Moreover, California boasts many higher educational institutions, which may not be the case in other states. Additionally, different states have varying population compositions, leading to discrepancies between the estimated and actual counts.

Additionally, sampling variability will also causes differences. If the data used in the estimation is a sample rather than a complete population census, then random sampling variability will affect the calculated ratio and the accuracy of the estimates.

Moreover, this method have bias. The Laplace ratio method works best when the relationship between the variable of interest and the population is consistent across different groups. However, California might does not represent other states due to factors, as discussed above, which will thus causes the results will be inaccurate.

In summary, assuming homogeneity in educational attainment across states when using ratio estimators can lead to differences between the estimates and actual numbers.

Appendix

A Instructions on How to Obtain the Data

To extract and download data from IPUMS USA, began by selecting “Get Data” and proceeded to “Select Sample.” Then deselected the “Default sample from each year” option and opted for the “2022 ACS” sample only. After confirming the selection, moved to the “Household” section, chose “Geographic,” and selected “STATEICP.” Similarly, under the “Person” section, selected “sex” and “EDUC.” Next, reviewed variable choices by clicking “View Cart” and then clicked “Create Data Extract.” Set the “Data Format” to “.csv” while keeping the “Data Structure” as “Rectangular.” After submitting the request, logged in the account, waited for the email notification that the extract was ready, and downloaded the file. Eventually, unzip the data using the command `gunzip usa_00002.csv.gz` to proceed with the analysis.

References

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ruggles, Steven, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. 2021. "IPUMS USA: Version 11.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/d010.v11.0.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wickham, Hadley, Evan Miller, and Danny Smith. 2023. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. https://CRAN.R-project.org/package=haven.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.