Estimating Doctoral Degree Holders Per State Using Ratio Estimators (2022 ACS Data)*

Jerry Xia   Peter Fan   Jason Yang

October 3, 2024

This paper utilizes the 2022 ACS IPUMS data to estimate the total number of respondents in each state based on the ratio of doctoral degree holders. Using California as a benchmark with 391,171 total respondents, we apply the Laplace ratio estimators approach to infer state-level totals and compare these estimates with the actual respondent numbers. The differences in estimates suggest possible variations in sample representation and other demographic factors.

Table of contents

[*] Code and data are available at: https://github.com/Jerryx2020/estimating_doctorates_per_state

1 Introduction

This paper applies the Laplace ratio estimator to the 2022 ACS IPUMS data (Ruggles et al. 2022) to estimate the total number of respondents per state based on the number of doctoral degree holders. The total number of respondents in California is known to be 391,171. Using the ratio of doctoral degree holders to total respondents in California, the same ratio is applied to estimate totals for other states. This approach helps understand the representation of doctoral degree holders and population variations across states.

The R software environment (R Core Team 2023) and the `tidyverse` (Hadley Wickham et al. 2019) were used for data manipulation, and the `ggplot2` package (H. Wickham 2016) was used to visualize the results. By comparing estimated and actual figures, the analysis explores discrepancies that may arise from demographic factors or sampling variability.

This paper is structured as follows: Section 2.1 provides an overview of the dataset and methodology, Section 2.4 presents key findings, and Section 3 discusses implications and potential refinements for the ratio estimators approach.

2 Data

2.1 Overview

The dataset used in this analysis is sourced from the 2022 American Community Survey (ACS) provided by IPUMS (Ruggles et al. 2022). The dataset contains individual-level data on respondents' education, gender, and state of residence, among other demographic factors.

The key variables used are:

- **STATEICP**: State identifier for each respondent.
- **EDUCD**: Educational attainment, where doctoral degree holders are identified by a specific code (116). Data were processed to focus on respondents with a doctoral degree (EDUCD code: 116) and grouped by state to obtain counts of doctoral degree holders per state.
- **PERWT**: Person-level weight to account for survey sampling.

2.2 Obtaining Data

The data we are using comes from IPUMS USA (Ruggles et al. 2022). To begin, you must first register for an account and log in. Then, navigate to the SELECT DATA page to locate the samples and variables you need. Click on the blue SELECT SAMPLE tab, select the 2022 ACS only, and SUBMIT SAMPLE SELECTIONS. After choosing the sample, proceed to SELECT HARMONIZED VARIABLES. We need three specific variables: HOUSEHOLD

-> GEOGRAPHIC -> STATEICP, PERSON -> DEMOGRAPHIC -> SEX, and PERSON -> EDUCATION -> EDUC. Once these are selected, click the VIEW CART tab and create a data extract. Change the data format to .csv and submit the extract. When it's ready, go to the MY DATA section, and click the green DOWNLOAD .CSV button. Finally, move the downloaded .csv file to the reflection folder.

2.3 Methodology

The ratio estimators approach is a statistical technique used to estimate population totals or means based on known ratios derived from a sample. This method involves calculating the ratio of a specific characteristic (e.g., the number of doctoral degree holders) to the total population for a known subset (e.g., California). This ratio is then applied to other subsets to estimate totals, assuming that similar relationships hold across the entire population. It is particularly useful when the exact population size is unknown, but a sample provides proportional relationships that can be generalized. The ratio of doctoral degree holders to the total number of respondents in California (391,171) serves as the basis for calculating estimated totals in other states.

The steps involved are: 1. Filtering the dataset to focus on doctoral degree holders using the `EDUCD` variable. 2. Computing the ratio: The ratio of doctoral degree holders in California is used to estimate totals for other states. 3. Comparing estimates: The estimated totals are compared with actual respondent numbers to highlight differences.

2.4 Results

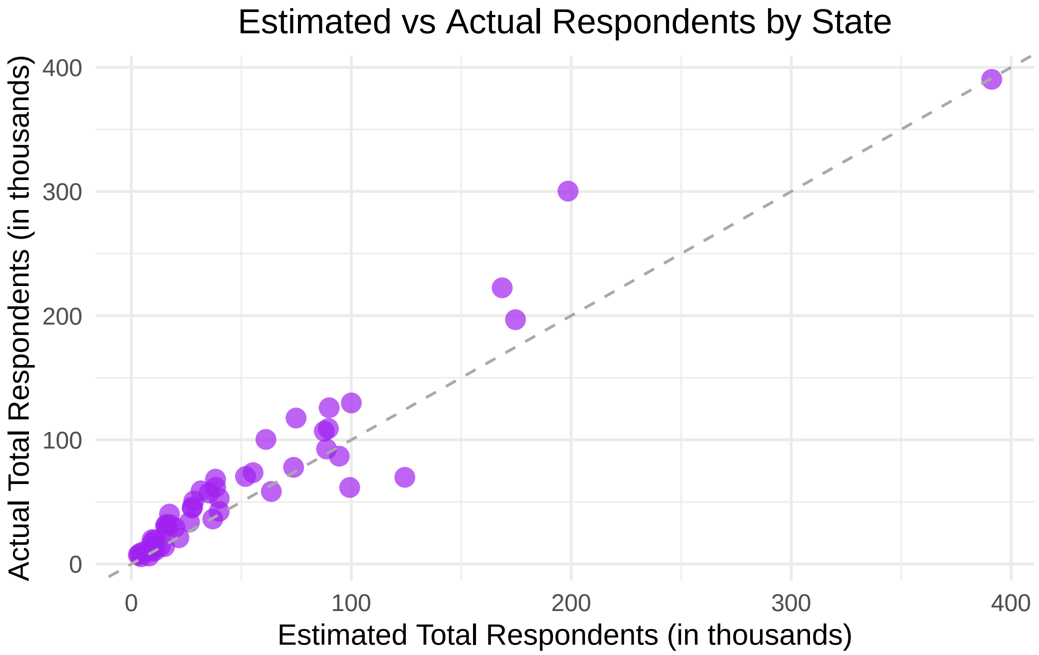

Figure 1: Estimated vs. Actual Respondents by State

The plot above compares the estimated total respondents per state with the actual respondent numbers. The dashed red line represents a perfect match between estimates and actuals. Points that deviate from this line suggest differences in the ratio of doctoral degree holders to the general population across states. It is scaled using the `ggplot2` package (H. Wickham 2016) `scales`.

3 Discussion

The results indicate that some states show larger discrepancies between the estimated and actual total respondents. These differences likely arise from variations in the proportion of doctoral degree holders across states. In some states, demographic factors such as urbanization, access to higher education, and population age distributions may result in over- or under-representation of doctoral degree holders.

Additionally, sampling variability within the ACS data may contribute to the observed differences. States with smaller populations or fewer respondents with doctoral degrees may have less accurate estimates due to larger sampling error.

Future refinements to the ratio estimators approach could involve incorporating additional demographic factors or using a more sophisticated weighting scheme to account for regional differences.

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/.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2022. *IPUMS: 2022 American Community Survey*. https://usa.ipums.org/usa/.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.