# Data Collection, Ratio Estimation, and Discrepancies: Analyzing ACS Respondent Counts with IPUMS USA*

Yi Tang Jin Zhang Siyuan Lu

October 3, 2024

The number of respondents in each state with a doctorate as their highest level of education is examined in this document using data from the 2022 ACS IPUMS. Using data from California, we use the ratio estimators approach to estimate the total number of respondents in each state.

Table 1: Counts of Respondents with Doctoral Degrees by State

| State (ICP Code) | Doctoral Degree Count |
|---|---|
| 1 | 600 |
| 2 | 165 |
| 3 | 2014 |
| 4 | 244 |
| 5 | 177 |
| 6 | 131 |
| 11 | 152 |
| 12 | 1438 |
| 13 | 2829 |
| 14 | 1620 |
| 21 | 1457 |
| 22 | 620 |
| 23 | 991 |
| 24 | 1213 |
| 25 | 513 |
| 31 | 258 |
| 32 | 321 |

*Code and data are available at: https://github.com/YiTang2/IPUMS-USA-data.git

| State (ICP Code) | Doctoral Degree Count |
|---|---|
| 33 | 572 |
| 34 | 621 |
| 35 | 153 |
| 36 | 60 |
| 37 | 71 |
| 40 | 1531 |
| 41 | 460 |
| 42 | 251 |
| 43 | 2731 |
| 44 | 1451 |
| 45 | 450 |
| 46 | 263 |
| 47 | 1421 |
| 48 | 647 |
| 49 | 3216 |
| 51 | 448 |
| 52 | 1608 |
| 53 | 281 |
| 54 | 841 |
| 56 | 159 |
| 61 | 896 |
| 62 | 1031 |
| 63 | 175 |
| 64 | 113 |
| 65 | 282 |
| 66 | 350 |
| 67 | 428 |
| 68 | 72 |
| 71 | 6336 |
| 72 | 647 |
| 73 | 1195 |
| 81 | 51 |
| 82 | 214 |
| 98 | 311 |

# 1 Introduction

We uses R packages (R Core Team 2023) to clean and analyze the dataset, including libraries from haven (Wickham, Miller, and Smith 2023), tidyverse (Wickham et al. 2019) and labelled

(Larmarange 2024). The data we used is from IPUMS (Ruggles et al. 2021).

## 2 A brief overview of the ratio estimators approach

The ratio estimator is a method used in survey sampling to improve estimation accuracy by leveraging a known relationship between two variables. This method calculates the ratio of a particular attribute to the total population for a known subset. The ratio is then applied to other subsets to approximate totals, assuming similar correlations exist across the entire population. It is especially helpful when the precise population size is unknown but a sample yields proportional connections.

## 3 Estimates and the Actual Number of Respondents

Table 2: Comparison of Actual vs. Estimated Respondent Counts by State

| State (ICP Code) | Actual Respondent Count | Estimated Respondent Count |
|---|---|---|
| 1 | 37369 | 37042.71 |
| 2 | 14523 | 10186.74 |
| 3 | 73077 | 124340.02 |
| 4 | 14077 | 15064.03 |
| 5 | 10401 | 10927.60 |
| 6 | 6860 | 8087.66 |
| 11 | 9641 | 9384.15 |
| 12 | 93166 | 88779.02 |
| 13 | 203891 | 174656.37 |
| 14 | 132605 | 100015.31 |
| 21 | 128046 | 89952.04 |
| 22 | 69843 | 38277.47 |
| 23 | 101512 | 61182.21 |
| 24 | 120666 | 74888.01 |
| 25 | 61967 | 31671.52 |
| 31 | 33586 | 15928.36 |
| 32 | 29940 | 19817.85 |
| 33 | 58984 | 35314.05 |
| 34 | 64551 | 38339.20 |
| 35 | 19989 | 9445.89 |
| 36 | 8107 | 3704.27 |
| 37 | 9296 | 4383.39 |
| 40 | 88761 | 94520.64 |

| State (ICP Code) | Actual Respondent Count | Estimated Respondent Count |
|---|---|---|
| 41 | 51580 | 28399.41 |
| 42 | 31288 | 15496.20 |
| 43 | 217799 | 168606.06 |
| 44 | 109349 | 89581.62 |
| 45 | 45040 | 27782.03 |
| 46 | 29796 | 16237.05 |
| 47 | 109230 | 87729.48 |
| 48 | 54651 | 39944.39 |
| 49 | 292919 | 198548.92 |
| 51 | 46605 | 27658.56 |
| 52 | 62442 | 99274.46 |
| 53 | 39445 | 17348.34 |
| 54 | 72374 | 51921.53 |
| 56 | 18135 | 9816.32 |
| 61 | 74153 | 55317.11 |
| 62 | 59841 | 63651.72 |
| 63 | 19884 | 10804.12 |
| 64 | 11116 | 6976.38 |
| 65 | 30749 | 17410.07 |
| 66 | 20243 | 21608.25 |
| 67 | 35537 | 26423.80 |
| 68 | 5962 | 4445.12 |
| 71 | 391171 | 391171.00 |
| 72 | 43708 | 39944.39 |
| 73 | 80818 | 73776.73 |
| 81 | 6972 | 3148.63 |
| 82 | 14995 | 13211.90 |
| 98 | 6718 | 19200.47 |

We calculates and compares the actual and estimated total respondent counts for each state using a ratio estimation method. It assumes the ratio of doctoral degree holders to total respondents in California applies to other states. The actual counts are derived from the dataset, while the estimated counts are based on this ratio. The final table highlights discrepancies between actual and estimated counts, helping evaluate the accuracy and limitations of the ratio estimation approach.

## 4 Explanation of why they are different

The estimated total number of respondents in each state using the ratio estimator can differ from the actual count due to several factors:

1. Assumption of Similarity: The ratio estimator makes the assumption that the percentage of Californians with doctorates is typical of other states; however, due to a variety of reasons, including economic, demographic, and educational infrastructure, there are considerable variations in educational attainment.
2. Sampling Variability: If based on a sample, random variability can impact the ratio and estimation accuracy.
3. Non-Uniform Distribution: Educational attainment isn't evenly distributed across the U.S., so any areas' ratio may not apply to other states.
4. Bias: When relationships are constant over all domains, the ratio technique performs well. The estimations will be skewed if the ratio is impacted by unobserved factors.

These factors explain why using the ratio estimator across diverse states often leads to differences from actual numbers.

# Appendix

## 5 Instructions on obtaining the data

To collect data from IPUMS USA, we first navigated to the IPUMS website and selected "IPUMS USA." We then clicked on "Get Data" and chose the "2022 ACS" sample under the "SELECT SAMPLE" section. To gather state-level data, we selected "HOUSEHOLD" followed by "GEOGRAPHIC" and added "STATEICP" to the cart. For individual-level data, we went to the "PERSON" section and added "EDUC" to the cart. After reviewing our selections by clicking "VIEW CART," we proceeded to "CREATE DATA EXTRACT." We changed the "DATA FORMAT" to ".csv" and clicked "SUBMIT EXTRACT." After logging in with the account, we received an email when the extract was ready for download. Finally, we downloaded the file to use in RStudio.

## References


Larmarange, Joseph. 2024. *Labelled: Manipulating Labelled Data*. https://CRAN.R-project.org/package=labelled.

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, 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, 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.

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