# A Case Study from 2022 IPUMS USA*

**Using the ratio estimation method is effective for estimating state populations, but significant differences exist between states.**

Jianing Li Xizi Sun Yawen Tan Shanjie Jiao Xinqi Yue
Duanyi Su

October 3, 2024

## Data tools

The dataset was analyzed using R (R Core Team 2023) and downloaded using the R package, Tidyverse (Wickham et al. 2019), Knitr (Xie 2014). Data are extract from Ipums_usa (IPUMS 2024).

## The original data

Table 1: Doctoral Degree Holders by State

| State Code | Doctoral Degree Holders |
|---:|---:|
| 1 | 600 |
| 2 | 165 |
| 3 | 2014 |
| 4 | 244 |
| 5 | 177 |
| 6 | 131 |
| 11 | 152 |
| 12 | 1438 |

*Code and data are available at: https://github.com/LilianS77/IPUMS_CA_EDUCD

## Overview of the Ratio Estimators Method for Population Estimation

The ratio estimators approach is a statistical technique used to estimate a total population based on a known ratio between two variables from a subset of the population. In my analysis, I first calculated the ratio of doctoral degree holders to the total number of respondents in California. This ratio provides a baseline to estimate the total number of respondents in other states by applying the same proportional relationship.

Once this ratio is determined, it is multiplied by the number of doctoral degree holders in each state to estimate the total population for that state. The underlying assumption here is that the relationship between doctoral degree holders and the overall population is similar across different states.

This method is useful when working with partial data, as it allows for generalization from a known sample to estimate unknown totals. However, it's important to note that the accuracy of this method depends on how well the ratio holds across different regions. Discrepancies between estimated and actual figures could arise due to variations in the distribution of educational attainment across states.

Table 2: Estimating Total Population Using Ratio and Merging with Actual Counts

| State Code | Estimated Total | Actual Total |
|---:|---:|---:|
| 1 | 37043 | 37369 |
| 2 | 10187 | 14523 |
| 3 | 124340 | 73077 |
| 4 | 15064 | 14077 |
| 5 | 10928 | 10401 |
| 6 | 8088 | 6860 |
| 11 | 9384 | 9641 |
| 12 | 88779 | 93166 |
| 13 | 174656 | 203891 |
| 14 | 100015 | 132605 |

## Reasons for Differences Between Estimated and Actual Values

The discrepancies between estimates obtained from the Laplace ratio estimator and actual values arise from several factors. One key issue is the representativeness of the sample data. If the sample does not accurately reflect the overall population, the estimates may deviate from the actual numbers. Additionally, foundational data may contain inaccuracies, such as incomplete data collection in certain regions or errors in the recorded values, which can distort the results.

The assumption that the ratio of doctoral degree holders to the total population is consistent across states may not hold true, given that different states have unique educational systems, economic conditions, and access to resources. Factors like population density, local policies, and economic circumstances can influence educational attainment and response rates in surveys, leading to variations in the data.

Moreover, sampling bias can occur due to differences in data sources, and the timing of data collection may not align perfectly across states, resulting in discrepancies in the statistics. Finally, statistical errors, such as rounding inaccuracies and the inherent limitations of estimation methods, can contribute to the differences between the estimated and actual values.

# Appendix

## How to Extract 2022 ACS Data from IPUMS

To obtain data from IPUMS, we start by navigating to the IPUMS USA section and clicking on Get Data. Next, we go to the Select Sample section, where we uncheck the "Default sample from each year" option and instead select 2022 ACS. After selecting our sample, we proceed to add variables of interest. For state-level data, we go to Household > Geographic and add STATEICP to our cart by clicking the plus icon next to it. For individual-level data, we might add variables from the Person section. For example, under Demographic, we could include variables like AGE, and under Person, we could add SEX and EDUCD (education attainment). Once our variables are selected, we click View Cart, then proceed by clicking Create Data Extract. At this point, we review our selections, change the Data Format to CSV, and submit our extract for processing. Then we saved it locally as usa_00002.csv.

# References

IPUMS. 2024. "IPUMS USA." https://usa.ipums.org/usa/#:~:text=IPUMS%20USA%20collects,%20preserves%20and%20harmonizes.

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, 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.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.