Estimating Doctoral Degree Holders Across U.S. States*

Using California Doctoral Degree Literacy Rates as a Reference for State-Level Estimations

Aviral Bhardwaj, Arshh Relan, Akshat Aneja, Yuxin Sun, Harsh Pareek

October 3, 2024

This report employs IPUMS data to estimate the number of doctoral degree holders in each U.S. state by using California's ratio of doctoral degree holders to total respondents as a benchmark. The analysis reveals significant variations in the estimated and actual counts of doctoral degree holders across states, highlighting discrepancies in educational attainment. This study highlights the importance of understanding educational distributions, as they influence workforce qualifications and inform policy decisions.

Instructions to download IPUMS data

1. Go to the IPUMS website
2. Click on "Select Samples" to choose the samples we want to download
3. Unselect "Default sample from each year"
4. Choose "2022 ACS"
5. Click on "Submit Sample Selections"
6. Navigate to the section called "Select Harmonized Variables"
7. Hover over "Household" and click on "Geographic"
8. Select "STATEICP" by clicking on the "+" icon.
9. Hover over "Person" and click on "Demographics"
10. Select "SEX" by clicking on the "+" icon.
11. Hover over "Person" and click on "Education"
12. Select "EDUC" by clicking on the "+" icon.
13. Click on "View Cart" and then "Create Data Extract"
14. Choose the format "CSV" under "Data Format"

*Code and data are available at: https://github.com/Aviral-03/IPUMS-Education-Doctorate

15. Click on “Submit Extract Request”
16. Create an account or log in to download the data

Data

The raw data was sourced from the IPUMS data (Ruggles et al. 2024) package. Three data points were added from year 2022: Sex of Respondants, Education background and State Codes. The data, provided in CSV formats, was cleaned and analyzed using R (R Core Team 2024) programming language. Other R packages used include `tidyverse` (Wickham et al. 2019), `styler` (Müller and Walthert 2024), and `dplyr` (Wickham et al. 2023) for creating tables. The `ggplot2` (Wickham 2016) and `kableExtra` (Zhu 2024) were used for data visualization and table formatting.

This analysis uses the IPUMS data to estimate the number of respondents with doctoral degrees in each state of the U.S. The data includes information on educational attainment, state of residence, and other demographic variables for a sample of respondents and sex of the respondents (Ruggles et al. 2024).

Table 1: Comparison of Respondents with Doctoral Degrees by State

State	Doctoral Degree Holders	Actual Total Respondents	Estimated Total Respondents	Difference
Massachusetts	2014	73077	124340.02	51263.02
Maryland	1608	62442	99274.46	36832.46
District of Columbia	311	6718	19200.47	12482.47
Virginia	1531	88761	94520.64	5759.64
Colorado	1031	59841	63651.72	3810.72
New Mexico	350	20243	21608.25	1365.25
Vermont	131	6860	8087.66	1227.66
New Hampshire	244	14077	15064.03	987.03
Rhode Island	177	10401	10927.60	526.60
California	6336	391171	391171.00	0.00

Discussion

The differences between the estimated and actual number of respondents in each state can arise from the assumption that the ratio of doctoral degree holders to total respondents in California is representative of all states. However, this assumption doesn't account for state-specific variations in educated population or socioeconomic factors.

California has unique characteristics, such as a large population, diverse industries, and numerous research institutions, which may lead to a higher concentration of doctoral degree holders compared to other states. States with smaller populations, fewer universities, or different economic structures might have lower proportions of doctoral degree holders.

Additionally, the distribution of educational attainment across the U.S. is not uniform. States with rural populations, different job markets, or less access to higher education may have lower rates of doctoral degrees, skewing the estimates.

The ratio estimator is a simplified approach, useful for generating rough estimates, but it overlooks local factors that significantly affect educational profiles across states, leading to deviations between the estimated and actual respondent numbers.

References

Müller, Kirill, and Lorenz Walthert. 2024. *styler: Non-Invasive Pretty Printing of R Code*. https://CRAN.R-project.org/package=styler.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rogers, and Megan Schouweiler. 2024. "IPUMS USA: Version 15.0." Dataset. IPUMS. https://doi.org/10.18128/D010.V15.0.

Wickham, Hadley. 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.

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

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.