California Education Attainment Census Data Analysis*

Rayan Awad Alim Emily Su Heyucheng Zhang
Maryam Ansari Prankit Bhardwai Luka Tosic

October 3, 2024

1 Data

We use the statistical programming language R (R Core Team 2023), IPUMS (@ Ruggles et al. 2022), dpylr (Wickham et al. 2023), here (Müller 2020), and tidyverse (Wickham et al. 2019). The data is taken from Ruggles et al. (2022).

1.1 Ratio Estimator Approach

How the ratio estimator approach works is that it estimates a population size based on a ratio of two means of information we know regarding the population. How we applied the ratio estimator is that we took the ratio of doctoral holders in California over the total number of respondents in California from our data. For each state we figured out the number of doctoral holders from our data and then divide this number with the ratio we obtained previously to get the estimated total number of respondents in each state.

*Code and data are available at: https://github.com/RayanAlim/US_Census_Education_Data_Analysis/

2 Results

3 Appendix

3.1 How to obtain IPUMS data

In order to obtain the data from IPUMS (@ Ruggles et al. 2022), first go to usa.ipums.org and then click on "Get Data" on the home page. Next click on "Select Samples", select only the 2022 ACS sample under the "USA SAMPLES" tab, and then "SUBMIT SAMPLE SELECTIONS". Then under "SELECT HARMONIZED VARIABLES" select "GEOGRAPHIC" under the "HOUSEHOLD" dropdown and choose the STATEICP variable. Under the PERSON dropdown, select EDUCATION and then the EDUC variable. After all this, click on VIEW CART on the top right and on the DATA CART Page click on "create data extract" and then on the Extract Request page make sure the data format is in the csv format. After submitting the extract on the Extract Request, create an account or log into IPUMS USA and wait for the extract to be finished and download the CSV.

References

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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. 2022. "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, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.