Estimating state populations based on California's doctorate holder ratio with ratio estimation

Andrew Goh, Yisu Hou, Liam Wall

Data was obtained from IPUMS 2022 ACS (Ruggles et al. 2024). R (R Core Team 2023), tidyverse (Wickham et al. 2019), and dplyr (Wickham et al. 2023) were used to analyse and modify the data. Styler (Müller and Walthert 2023) was used to style the code files.

How to obtain the data: data was obtained from IPUMS 2022 ACS - first, we created an account on IPUMS. Then, we selected only the 2022 ACS dataset, taking the variables sex (SEX under person-demographic), state (STATEICP under household-geographic), and highest education taken (EDUC under person-education). Once we selected the correct variables, we then submitted the data set for review, downloaded the extract (in .csv format), then used the codebook for reference to match integer values with the data represented.

The ratio estimators approach involves using the ratio of two means for a particular sample to extrapolate an estimate for other, similar, populations. For this particular example, we used the dataset to find the ratio between the total number of doctorates in California and the total number of respondents in California, then used that ratio to estimate the total number of respondents in each other state by applying the ratio to the total number of doctorates in said states.

Table 1 shows the estimates (via the ratio estimation method) and actual number of respondents.

Table 1: Estimated Total Respondents (via ratio estimation method) vs Actual Total Respondents by State

State	Actual_Total_Respondents	Estimated_Total_Respondents
Connecticut	37369	37043
Maine	14523	10187
Massachusetts	73077	124340
New Hampshire	14077	15064

State	Actual_Total_Respondents	Estimated_Total_Respondents
Rhode Island	10401	10928
Vermont	6860	8088
Delaware	9641	9384
New Jersey	93166	88779
New York	203891	174656
Pennsylvania	132605	100015
Illinois	128046	89952
Indiana	69843	38277
Michigan	101512	61182
Ohio	120666	74888
Wisconsin	61967	31672
Iowa	33586	15928
Kansas	29940	19818
Minnesota	58984	35314
Missouri	64551	38339
Nebraska	19989	9446
North Dakota	8107	3704
South Dakota	9296	4383
Virginia	88761	94521
Alabama	51580	28399
Arkansas	31288	15496
Florida	217799	168606
Georgia	109349	89582
Louisiana	45040	27782
Mississippi	29796	16237
North Carolina	109230	87729
South Carolina	54651	39944
Texas	292919	198549
Kentucky	46605	27659
Maryland	62442	99274
Oklahoma	39445	17348
Tennessee	72374	51922
West Virginia	18135	9816
Arizona	74153	55317
Colorado	59841	63652
Idaho	19884	10804
Montana	11116	6976
Nevada	30749	17410
New Mexico	20243	21608
Utah	35537	26424
Wyoming	5962	4445

State	Actual_Total_Respondents	Estimated_Total_Respondents
California	391171	391171
Oregon	43708	39944
Washington	80818	73777
Alaska	6972	3149
Hawaii	14995	13212
District of Columbia	6718	19200

We notice that the difference in estimated vs actual total respondents is only reasonably accurate (within 1 million people difference) for 4 of the 51 geographical areas listed in the data. Potential reasons for this include how the public education system in California is stronger than that of most other states, leading to a higher percentage of doctorate degree holders than that of other regions in the USA, potentially leading to inaccurate total respondents estimates. California's high average GDP might also lead to a wealthier demographic in total, which in turn correlates with a larger percentage of doctorate holders in the population as a whole.

References

Müller, Kirill, and Lorenz Walthert. 2023. *Styler: Non-Invasive Pretty Printing of r Code*. https://github.com/r-lib/styler.

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, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. 2024. *IPUMS USA: Version 15.0*. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.

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://dplyr.tidyverse.org.