California's Education Attainment Census Data Analysis*

Jinyan Wei Amy Jin

November 21, 2024

1 Data

The dataset was analyzed using R (R Core Team 2023) and downloaded using the R package, dpylr (Wickham et al. 2023) Knitr (Xie 2014) and here (Müller 2020). The data we used is taken from IPUMS (Ruggles et al. 2022).

We are interested in "how many respondents were there in each state (STATEICP) that had a doctoral degree as their highest educational attainment (EDUC)?"

Table 1: Counts of Respondents with a Doctoral

State	Number of Doctoral Degree Holders
connecticut	600
maine	165
massachusetts	2,014
new hampshire	244
rhode island	177
vermont	131
delaware	152
new jersey	1,438
new york	2,829
pennsylvania	1,620
illinois	1,457
indiana	620

*A GitHub Repository containing all data, R code, and other files used in this investigation is located here: https://github.com/jeno0403/US-Census-Education-Data-Analysis

michigan	991
ohio	1,213
wisconsin	513
iowa	258
kansas	321
minnesota	572
missouri	621
nebraska	153
north dakota	60
south dakota	71
virginia	1,531
alabama	460
arkansas	251
florida	2,731
georgia	1,451
louisiana	450
mississippi	263
north carolina	1,421
south carolina	647
texas	3,216
kentucky	448
maryland	1,608
oklahoma	281
tennessee	841
west virginia	159
arizona	896
colorado	1,031
idaho	175
montana	113
nevada	282
new mexico	350
utah	428
wyoming	72
california	6,336
oregon	647
washington	1,195
alaska	51
hawaii	214
district of columbia	311

1.1 Ratio Estimator Approach

The ratio estimator approach is a statistical method used to estimate an unknown population total or average by leveraging the relationship between two correlated variables. It is particularly useful in surveys and studies where one variable is easier to measure than the target variable. By utilizing the proportion between these two variables, the ratio estimator allows us to extend estimates to the entire population.

In our case, we aim to estimate the number of respondents with a doctoral degree in every state using the ratio estimator, given our limited information on California.

Detailed Steps in the Ratio Estimator Approach:

1. First, we determine the ratio between two known quantities for a subset of the population. For example, the ratio of respondents with doctoral degrees to the total number of respondents in California can be expressed as: $R = \frac{Respondents with Doctoral Degrees in California}{Total Respondents in California}$
2. After calculating the ratio for California, we assume that this ratio is representative of the same relationship in other states. This ratio is then used to estimate the total number of respondents in each state by applying it to the number of respondents with doctoral degrees in those states. The formula is as follows: $\hat{N}_{state} = \frac{Respondents with Doctoral Degrees in State}{R}$

2 Results

Table 2: Comparison of Estimated vs. Actual Respondents by State Services

State	Actual Total Respondents	Estimated Total Respondents
connecticut	37,369	37,043
maine	14,523	10,187
massachusetts	73,077	124,340
new hampshire	14,077	15,064
rhode island	10,401	10,928
vermont	6,860	8,088
delaware	9,641	9,384
new jersey	93,166	88,779
new york	203,891	174,656
pennsylvania	132,605	100,015
illinois	128,046	89,952
indiana	69,843	38,277
michigan	101,512	61,182
ohio	120,666	74,888

wisconsin	61,967	31,672
iowa	33,586	15,928
kansas	29,940	19,818
minnesota	58,984	35,314
missouri	64,551	38,339
nebraska	19,989	9,446
north dakota	8,107	3,704
south dakota	9,296	4,383
virginia	88,761	94,521
alabama	51,580	28,399
arkansas	31,288	15,496
florida	217,799	168,606
georgia	109,349	89,582
louisiana	45,040	27,782
mississippi	29,796	16,237
north carolina	109,230	87,729
south carolina	54,651	39,944
texas	292,919	198,549
kentucky	46,605	27,659
maryland	62,442	99,274
oklahoma	39,445	17,348
tennessee	72,374	51,922
west virginia	18,135	9,816
arizona	74,153	55,317
colorado	59,841	63,652
idaho	19,884	10,804
montana	11,116	6,976
nevada	30,749	17,410
new mexico	20,243	21,608
utah	35,537	26,424
wyoming	5,962	4,445
california	391,171	391,171
oregon	43,708	39,944
washington	80,818	73,777
alaska	6,972	3,149
hawaii	14,995	13,212
district of columbia	6,718	19,200

The estimates and the actual number of respondents in Table 2 are different because the ratio estimators approach doesn't consider different factors like environment, socio-economic status

of respondents, etc. that could impact a respondent's highest educational attainment to be a doctoral degree or not. The ratio estimators approach assumes that all states have the same factors as California that impacts a respondent's highest educational attainment.

3 Discussion

3.1 Limitations of Assumptions in Ratio Estimation

The ratio estimator assumes that the proportion of respondents with doctoral degrees in California is representative of the proportion in other states. However, this assumption may not hold due to significant differences in educational attainment across states, influenced by varying demographics, economic opportunities, and educational infrastructure. Additionally, the distribution of educational attainment is non-uniform across the United States. Factors like regional policies, cultural differences, and access to higher education contribute to this variation, making the California ratio less applicable elsewhere. California's status as the state with the highest GDP in 2022 further complicates the assumption of consistency, as it breaks the premise of Laplace's method that specific characteristics in a sample are uniformly distributed across different subsets of the population.

3.2 Variation in education attainment across states

The ratio estimator only takes the proportion of respondents with a doctoral degree in California. The rest of the states are calculated under the assumption that the actual higher education rate across all states is the same (or similar), but this could be too big of an assumption. In fact, the education attainment degree may vary a lot. There are both personal and societal factors that could contribute to this difference. For instance, access to higher education, state policy, and local funding are all important factors that could have a big impact.

3.3 Variation in Sample Size

Remember that Laplace's ratio estimator works best when dealing with a larger sample size. If the sample size of other states is too small, variance increases, and the estimator does not do as well as it could have. If the state has relatively few amount of doctoral degrees, this ratio is not going to be a good reflection of its states. Also, if the sample size is too small, the outliers effect will increase; extreme data will overly skew the estimator, thus making the wrong prediction.

3.4 Bias in the Ration

In the context of the Laplace ratio approach, bias in the ratio occurs when the relationship between the characteristic of interest (e.g., the number of doctoral degree holders) and the total population is not consistent across different regions or units. Since holding a doctoral degree is not a common characteristic in the general population, applying a ratio based on one subset (e.g., California) to other subsets (e.g., other states) can introduce bias if unobserved factors affect the ratio in those other regions. If the ratio of doctoral degree holders to the total population in California is not representative of other states due to differences such as educational infrastructure, economic factors, or demographics, the estimates produced using the Laplace ratio approach will be biased and may overestimate or underestimate the actual totals.

These reasons indicate that the assumption of homogeneity used in ratio estimators often leads to differences when applied to diverse populations such as different states in the US.

4 Appendix

4.1 How to obtain IPUMS data

To obtain the data from IPUMS (Ruggles et al. 2022), we started by selecting "IPUMS USA" on the IPUMS website, then clicked "Get Data" and selected "2022 ACS" under "SELECT SAMPLE." We specified state-level data by selecting "HOUSEHOLD" > "GEOGRAPHIC" and added "STATEICP" to our cart by clicking the plus. We went to "PERSON" and added "EDUC" to our cart for individual-level data. We clicked "VIEW CART" and then "CREATE DATA EXTRACT." We changed the "DATA FORMAT" to ".dta." Finally, we want to include a descriptive name for our extract, for instance, "2024-10-03: State, education", which specifies the date we made the extract and what is in the extract. After that, we can click "SUBMIT EXTRACT". After logging in or creating an account, we received an email when the extract was ready. We then downloaded and saved it locally (e.g., "usa_00004.dta") for use in `R`.

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

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. http://www.crcpress.com/product/isbn/9781466561595.