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.

2 Results

---

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

Table 1: Comparison of Estimated vs. Actual Respondents by State

State ICP Code	Doctoral Degree Holders	Estimated Total Respondents	Actual Total Respondents
1	600	37042.708	37369
2	165	10186.745	14523
3	2014	124340.024	73077
4	244	15064.035	14077
5	177	10927.599	10401
6	131	8087.658	6860
11	152	9384.153	9641
12	1438	88779.024	93166
13	2829	174656.370	203891
14	1620	100015.312	132605
21	1457	89952.043	128046
22	620	38277.465	69843
23	991	61182.207	101512
24	1213	74888.009	120666
25	513	31671.516	61967
31	258	15928.365	33586
32	321	19817.849	29940
33	572	35314.049	58984
34	621	38339.203	64551
35	153	9445.891	19989
36	60	3704.271	8107
37	71	4383.387	9296
40	1531	94520.644	88761
41	460	28399.410	51580
42	251	15496.200	31288
43	2731	168606.061	217799
44	1451	89581.616	109349
45	450	27782.031	45040
46	263	16237.054	29796
47	1421	87729.481	109230
48	647	39944.387	54651
49	3216	198548.917	292919
51	448	27658.556	46605
52	1608	99274.458	62442
53	281	17348.335	39445
54	841	51921.530	72374
56	159	9816.318	18135
61	896	55317.111	74153

State ICP Code	Doctoral Degree Holders	Estimated Total Respondents	Actual Total Respondents
62	1031	63651.720	59841
63	175	10804.123	19884
64	113	6976.377	11116
65	282	17410.073	30749
66	350	21608.247	20243
67	428	26423.799	35537
68	72	4445.125	5962
71	6336	391171.000	391171
72	647	39944.387	43708
73	1195	73776.727	80818
81	51	3148.630	6972
82	214	13211.899	14995
98	311	19200.470	6718

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.