California's Education Attainment Census Data Analysis*

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

October 4, 2024

1 Data

We used the statistical programming language R (R Core Team 2023), dpylr (Wickham et al. 2023), here (Müller 2020), and tidyverse (Wickham et al. 2019) to assist with cleaning and analyzing the data. The data we used is taken from IPUMS (Ruggles et al. 2022). We rendered our table using knitr (Xie 2014).

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 Services

State ICP Code	Doctoral Degree Holders	Estimated Total Respondents	Actual Total Respondents
1	600	37043	37369
2	165	10187	14523
3	2014	124340	73077
4	244	15064	14077
5	177	10928	10401
6	131	8088	6860
11	152	9384	9641
12	1438	88779	93166
13	2829	174656	203891
14	1620	100015	132605
21	1457	89952	128046
22	620	38277	69843
23	991	61182	101512
24	1213	74888	120666
25	513	31672	61967
31	258	15928	33586
32	321	19818	29940
33	572	35314	58984
34	621	38339	64551
35	153	9446	19989
36	60	3704	8107
37	71	4383	9296
40	1531	94521	88761
41	460	28399	51580
42	251	15496	31288
43	2731	168606	217799
44	1451	89582	109349
45	450	27782	45040
46	263	16237	29796
47	1421	87729	109230
48	647	39944	54651
49	3216	198549	292919
51	448	27659	46605
52	1608	99274	62442
53	281	17348	39445
54	841	51922	72374
56	159	9816	18135
61	896	55317	74153

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

State ICP Code	Doctoral Degree Holders	Estimated Total Respondents	Actual Total Respondents
62	1031	63652	59841
63	175	10804	19884
64	113	6976	11116
65	282	17410	30749
66	350	21608	20243
67	428	26424	35537
68	72	4445	5962
71	6336	391171	391171
72	647	39944	43708
73	1195	73777	80818
81	51	3149	6972
82	214	13212	14995
98	311	19200	6718

The estimates and the actual number of respondents in Table 1 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 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.

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.