Estimating the tital number of respondants based on Doctorate to Total ratio in California*

Shreya Sakura Noskor Zeyi Cai Mo (Molly) Zhou
Bingxu Li

November 21, 2024

We estimate the total number of respondants to the survey based on an estimate, where we assume that all states have the same doctorate to total ratio. This is done by calculating the ratio for california and applying that to all states. We see in most cases that the estimate is lower than the actual number, which implies that the california has a higher number of people who hold doctorates. This is important as we want to analyse which state has the most number of doctorates.

1 Introduction

In this analysis, we aim to estimate the total number of respondents based on ratio of doctorate-holding people in California applied to number of doctorate-holding persons in other U.S. states. We assume that all states have the same proportion of doctoral-degree holders relative to the total number of survey respondents as California. To do this, we first calculate the doctorate-to-total ratio for California and apply that ratio to each state in the dataset.Our estimand is the actual number of respondants of the survey carried out bu the U.S. government.

However, our findings indicate that in most states, the estimated number of doctoral-degree holders is lower than the actual number. This suggests that California has a higher proportion of respondents with doctoral degrees compared to other states. This insight is critical for understanding the distribution of doctoral education across the country, as it highlights the relative education levels of states and can inform further analysis of which states are leading in the number of individuals with doctoral degrees.

Through this methodology, we aim to better understand the educational landscape across the U.S. and identify states with the highest concentrations of highly educated individuals,

*Code and data are available at: https://github.com/NotSakura/RatioEstimator.git.

particularly those with doctoral qualifications. Our data section (Section 2) cites the dataset and the variables we used. Our result section outlines the actual estimated numbers we get. And lastly our Discussion section (Section 4) tells us about what the Result (Section 3) section means.

2 Data

The data used in the paper to investogate ratio estimators are collected from USA (2024). We used R (R Core Team 2023) to write the code and the data was cleaned using tidyverse (Wickham et al. 2019) and haven (Wickham, Miller, and Smith 2023)

2.1 Measurement

This data is collected by the US government in their census. The data is collected with surevy and such which is why it is a reliable data collection. It is of course no perfect as there could be missing data from the under-represented sample of the population but on average they are pretty accurate. The variables we explore is state (in order to filter by the state of California), educd (which tells us our education level).

2.2 Instructions on Extracting Data

We first go to the IPUMS american data website and then we click 'get data' https://usa.ipums.org/usa/. Next we click SELECT SAMPLES then we deselect everything there. We then click ASC 2022 to get the 2022 data. Then we check which values we want so we go to Households and pick Geographic, where we select STATEICP. And we go to Person and click education where we click EDUC. Once we finished that we click view cart to see our dataset (click CREATE DATA EXTRACT). We change the data format to .dta and we change the sample size to 500 after we click CUSTOMISE SAMPLE SIZES. Then we give and an appropriate detail fo extract that includes today's date and what columns we are going to be looking at. Then we download it (after making an account) and add it to our repository to read.

2.3 Brief Overview of The Ratio Estimators Approach

Referred to the textbook to follow the approach of finding the ratio estimators. we first filtered for the state of California. Then we counted how many actual respondents there were and how many of them have their doctorate. We then divide it to find the ratio and then we apply that to the number of doctorate person in each state.

3 Results

3.1 Ratio and Actual Number of Respondants

Table 1: The estimated, and actual repondants of each state, represented by state ICP

State ICP	Doctoral Respondents	Actual Respondents	Estimated Total Respondents
1	80	5,518	4,983
2	29	2,175	1,806
3	274	10,946	17,067
4	37	2,151	2,305
5	25	1,514	1,557
6	13	1,002	810
11	23	1,404	1,433
12	197	13,692	12,270
13	430	30,064	26,783
14	239	19,503	14,887
21	222	19,134	13,828
22	97	10,436	6,042
23	139	14,953	8,658
24	181	17,902	11,274
25	78	9,205	4,858
31	38	4,965	2,367
32	51	4,457	3,177
33	90	8,628	5,606
34	93	9,644	5,793
35	15	2,918	934
36	6	1,212	374
37	11	1,465	685
40	209	13,216	13,018
41	58	7,631	3,613
42	40	4,645	2,491
43	368	32,350	22,922
44	242	16,051	15,073
45	80	6,804	4,983
46	30	4,423	1,869
47	241	16,057	15,011
48	93	8,084	5,793
49	470	43,739	29,275
51	64	6,888	3,986
52	249	9,349	15,509

State ICP	Doctoral Respondents	Actual Respondents	Estimated Total Respondents
53	39	5,775	2,429
54	137	10,746	8,533
56	24	2,620	1,495
61	132	10,961	8,222
62	164	8,899	10,215
63	30	2,936	1,869
64	27	1,636	1,682
65	46	4,569	2,865
66	58	3,014	3,613
67	51	5,243	3,177
68	8	865	498
71	931	57,989	57,989
72	119	6,604	7,412
73	176	11,913	10,962
81	3	1,033	187
82	38	2,217	2,367
98	28	987	1,744

4 Discussion

California is home to many universities and research institutions, which may contribute to a higher number of doctoral degree holders relative to the population. This uniqueness can skew the ratio when applied to other states that do not share the same educational environment.

Also the ratio estimator assumes that the relationship between doctoral respondents and total respondents in California is same for all other states. This may not be true as in some states there may not be as much universities or the proportion of regular students to doctorate students could be different.

Because of the above, in Table 1 we see that some states have a lower estimate than the actual number of respondents, which is due to the fact that their ratio between people holding doctorates vs. all people is lower than the ratio of California.

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

USA, IPUMS. 2024. "IPUMS USA: United States Census Data." https://usa.ipums.org/usa/.

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, Evan Miller, and Danny Smith. 2023. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. https://CRAN.R-project.org/package=haven.