U.S.A. Doctoral Respondents Estimation Analysis ACS 2022*

Steven Li Tim Chen Xinxiang Gao John Zhang
Tommy Fu Sandy Yu

October 4, 2024

This paper provides an analysis of the total number of doctoral respondents in the 2022 American Census Survey (ACS) using data from IPUMS USA. The authors use a Laplace ratio estimation method, where the ratio of doctoral respondents to the total population in California is applied to other states to estimate their respondent counts. The study highlights a mean difference of 19.56% between the estimated and actual respondent counts, pointing to potential discrepancies. These discrepancies are attributed to variations in educational attainment across states, emphasizing the limitations of using a single ratio estimator.

*Code and data are available at: https://github.com/stevenli-uoft/US_Doctoral_Respondents_Analysis

1 Introduction

This paper outlines the number of doctoral respondent by state in 2022 American Census Survey and proceeds to estimate the total number of respondents using California's doctoral respondents count. The data used in this paper is collected from IPUMS USA (2022).

The remainder of this paper is structured as follows. Section 2 provides a sample look at the data. Section 3 discusses the LaPlace estimation methods. Section 4 presents the LaPlace estimation results. Section 5 dives into the explanation and reasoning behind differences.

The dataset was cleaned and processed using R (R Core Team 2023), with additional support from the tidyverse (Wickham et al. 2019) packages. The cleaning process involved removing any unnecessary variables, and calculating the LaPlace estimations.

2 Data

Table 1 is a sample of the downloaded data from IPUMS USA (2022), and the columns needed for our analysis.

Table 1: Sample Data

STATEICP	EDUCD	SEX
32	26	2
3	26	2
40	12	2
13	22	2
49	63	1

3 Brief Overview of the Ratio Estimators Approach

The ratio estimators approach, also known as the Laplace ratio estimator, is a statistical method used to estimate population parameters when only partial information is available. In this case, we're using it to estimate the total number of respondents in each state based on the known number of respondents with doctoral degrees.

The basic idea behind this approach is to use a known ratio from one population (in this case, California) and apply it to other populations to estimate their total size. The steps involved are:

1. Calculate the ratio of doctoral degree holders to total respondents in California. (Assume this ratio is constant across all states)

2. For each state, divide the number of doctoral degree holders by this ratio to estimate the total number of respondents.

This method relies on the assumption that the proportion of doctoral degree holders is relatively consistent across states, which may not always be true in practice.

4 Estimates and Actual Number of Respondents

Table 3 in the appendix presents the total doctoral count, total respondents, and estimated respondent count for every state. Table 2 shows the summary statistics of Table 3, presenting a mean difference of 19.56% between estimated and actual respondents.

Table 2: Laplace Estimation Summary Statistics

Mean Difference	Median Difference	Mean Percent Difference	Median Percent Difference
12785.06	10122	19.56	28.26

5 Explanation of Differences

Our estimates using the Laplace ratio estimator show some notable differences from the actual numbers of respondents in each state. Here are the key points to consider:

- **Magnitude of differences:** On average, our estimates differed from the actual numbers by about 12,785 respondents (mean difference), with a median difference of 10,122. This suggests that while some states had larger discrepancies, the typical difference was around 10,000 respondents.
- **Variation in education levels:** The primary reason for these differences is likely the variation in educational attainment across states. Our method assumed a constant ratio of doctoral degree holders to total population based on California's data. However, this ratio almost certainly varies between states due to differences in economic structures, presence of research institutions, and demographic compositions.

These findings highlight the limitations of applying a single ratio estimator across diverse populations and emphasize the need for more nuanced approaches when estimating population parameters across different regions.

6 Appendix

6.1 Instructions on how to obtain the data:

1. Go to https://usa.ipums.org/usa/
2. Create an account or log in
3. Select the 2022 ACS sample
4. Choose the following variables: STATEICP, EDUC, SEX
5. Submit the extract request
6. Download the data and save it as "usa_00001.csv" in a "data" folder in your project directory gunzip usa_00001.csv.gz
7. If you have problems opening the zip file:

    1. Open your terminal
    2. Navigate to the folder containing the zip file
    3. Paste gunzip usa_00001.csv.gz into the terminal, and click enter

8. Move the usa_00001.csv to the folder "data/"

6.2 Processed Data

Table 3: State Doctoral and Respondant Counts, and Estimates

STATEICP	Actual Doctoral Count	Total Respondent	Estimated Respondent Count	Difference	% Difference
71	6336	391171	391171	0	0.00
49	3216	292919	198549	94370	32.22
13	2829	203891	174656	29235	14.34
43	2731	217799	168606	49193	22.59
3	2014	73077	124340	-51263	-70.15
14	1620	132605	100015	32590	24.58
52	1608	62442	99274	-36832	-58.99
40	1531	88761	94521	-5760	-6.49
21	1457	128046	89952	38094	29.75
44	1451	109349	89582	19767	18.08
12	1438	93166	88779	4387	4.71
47	1421	109230	87729	21501	19.68
24	1213	120666	74888	45778	37.94
73	1195	80818	73777	7041	8.71
62	1031	59841	63652	-3811	-6.37
23	991	101512	61182	40330	39.73
61	896	74153	55317	18836	25.40

Table 3: State Doctoral and Respondant Counts, and Estimates

STATEICP	Actual Doctoral Count	Total Respondent	Estimated Respondent Count	Difference	% Difference
54	841	72374	51922	20452	28.26
48	647	54651	39944	14707	26.91
72	647	43708	39944	3764	8.61
34	621	64551	38339	26212	40.61
22	620	69843	38277	31566	45.20
1	600	37369	37043	326	0.87
33	572	58984	35314	23670	40.13
25	513	61967	31672	30295	48.89
41	460	51580	28399	23181	44.94
45	450	45040	27782	17258	38.32
51	448	46605	27659	18946	40.65
67	428	35537	26424	9113	25.64
66	350	20243	21608	-1365	-6.74
32	321	29940	19818	10122	33.81
98	311	6718	19200	-12482	-185.80
65	282	30749	17410	13339	43.38
53	281	39445	17348	22097	56.02
46	263	29796	16237	13559	45.51
31	258	33586	15928	17658	52.58
42	251	31288	15496	15792	50.47
4	244	14077	15064	-987	-7.01
82	214	14995	13212	1783	11.89
5	177	10401	10928	-527	-5.07
63	175	19884	10804	9080	45.66
2	165	14523	10187	4336	29.86
56	159	18135	9816	8319	45.87
35	153	19989	9446	10543	52.74
11	152	9641	9384	257	2.67
6	131	6860	8088	-1228	-17.90
64	113	11116	6976	4140	37.24
68	72	5962	4445	1517	25.44
37	71	9296	4383	4913	52.85
36	60	8107	3704	4403	54.31
81	51	6972	3149	3823	54.83

References

IPUMS USA, University of Minnesota. 2022. "IPUMS USA: Version 12.0 [Dataset]." https://usa.ipums.org/usa/.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.