Analyzing State-Level Respondent Estimates Using Ratio Estimators: A Case Study with 2022 ACS Data*

Yunkai Gu, Anqi Xu, Yitong Wang

November 21, 2024

This study analyzes state-level respondent estimates using data from the 2022 American Community Survey (ACS) accessed via IPUMS, and focuses on individuals with doctoral degrees and employs the ratio estimator approach to predict total respondents in each state, using California as a reference point. The results reveal discrepancies between estimates and actual figures, highlighting the limitations of assuming uniform ratios across states. These differences are attributed to variations in demographics, educational policies, and economic conditions. The analysis underscores the utility of statistical methods like ratio estimators for scalable estimates while emphasizing the importance of accounting for regional variations in such analyses.

Table of contents

---

*Code and data are available at: https://github.com/Anjojoo/State-Level-Respondent-Estimates.

1 Introduction

This paper utilizes the 2022 American Community Survey (ACS) data from IPUMS to explore educational attainment across states, focusing on respondents with doctoral degrees. The task involves determining the number of respondents with a doctoral degree in each state (STATEICP) and applying the ratio estimator approach of Laplace. Using California as a benchmark, where the total number of respondents across all education levels is known, the ratio of doctoral degree holders to total respondents is calculated. This ratio is then applied to other states to estimate their total respondent counts. The analysis includes a comparison of these estimates with actual state-level respondent numbers, highlighting differences and discussing potential causes for discrepancies.

The remainder of this paper is structured as follows:

Section 2 introduces the data obtaining process (Section 2.1), measurement (Section 2.2), as well as an overview of the ratio estimators approach (Section 2.3). Then, Section 3 presents the estimates and the actual number of respondents, and Section 4 discusses the results, and reasons of difference between the estimated result and the actual value.

2 Data

2.1 Data obtaining

We gather the data from IPUMS USA site, firstly we select "IPUMS USA" on the IPUMS, then clicked "Get Data", then click "SELECT SAMPLE" and only select "2022 ACS". We choose state level data by selecting "HOUSEHOLD", then choose "GEOGRAPHIC" and add "STATEICP" to cart. For individual level data, we directly search "EDUC" and add it to the cart. After that, we clicked "VIEW CART" , then click "CREATE DATA EXTRACT". We modify the it to csv form. We clicked "SUBMIT EXTRACT" and download it. The data will not be uploaded to github due to its large size, and the prohibition of IPUMs.

The analyses presented in this paper were conducted using R programming language [@citeR]. The `tidyverse` packages [@citetidyverse] were used in the process of data simulation, testing beforehand. After the original raw data was downloaded by using `tidyverse` package [@citetidyverse], data cleaning process was done by using `tidyverse` package [@citetidyverse]. We use `testthat` package [@citetest] to develop the test for structure and format of simulation and analysis data. Tables were constructed with `knitr` package [@citeknitr] and `tibble` package [@citetibble].

Table 1 shows the number of respondents that had a doctoral degree as their highest educational attainment (EDUC) in each state (STATEICP).

Table 1: Respondents with a doctoral degree as their highest educational attainment in each state

State	Doctoral Counts
1	600
2	165
3	2014
4	244
5	177
6	131
11	152
12	1438
13	2829
14	1620
21	1457
22	620
23	991
24	1213
25	513
31	258
32	321
33	572
34	621
35	153
36	60
37	71
40	1531
41	460
42	251
43	2731
44	1451
45	450
46	263
47	1421
48	647
49	3216
51	448
52	1608
53	281

Table 1: Respondents with a doctoral degree as their highest educational attainment in each state

State	Doctoral Counts
54	841
56	159
61	896
62	1031
63	175
64	113
65	282
66	350
67	428
68	72
71	6336
72	647
73	1195
81	51
82	214
98	311

2.2 Measurement

This paper uses IPUMS to access 2022 American Community Survey (ACS), focusing on the real-world phenomenon of educational attainment — specifically, individuals obtaining a doctoral degree. As society has been paying arising attention to a specific person's education progress, we find this widespread social phenomenon a meaningful topic to make investigations on. We used the raw dataset from IPUMS USA, which provides a large database containing data collected from surveys or census activities by U.S. Census and American Community Survey from 1850 to the present. [@ipums-us] The raw data was downloaded from the website and stored in a structured table with columns of STATEICP (indicating the state information of each respondent), EDUC (indicating highest educational attainment of each respondent, i.e. a doctoral degree, which is the topic we have been working on) and other columns presenting various pieces of information gathered from the respondent. Then, we cleaned the raw dataset, removing incomplete and unnecessary columns and rows to make our analysis clearer. In particular, we selected the EDUC column and STATEICP column to keep valuable information and processed the data to show the number of respondents in each state that had a doctoral degree as their highest educational attainment. Steps went on to the estimation approach, where we used Laplace ratio estimator to estimate the total number of respondents in each state, given the number of respondents (391,171) in California. This is a statistical method to obtain counts of respondents without acknowledging every piece of information, which is

convenient and easy for computing following statistical analysis. Although the method could be helpful, it may also lead to biases. This is because the Laplace estimating method is based on the assumption that the proportion of doctoral degree holders is similar across states, which could not be held in real-life situations. Therefore, the differences between real-world data (actual survey responses) and statistical estimates as inferred from California's data was calculated to investigate whether the ratio provides a good estimate and to consider the underlying reason explaining the differences.

2.3 Overview of the ratio estimators approach

The ratio estimator is a method used to improve the accuracy of estimates for a population parameter when there is an auxiliary variable related to the variable of interest. In this case, the objective is to estimate the total number of respondents in each state in the 2022 ACS dataset, given the known number of respondents with doctoral degrees in each state and the California ratio.

With the given total number of respondents in California across all education levels and the number of respondents in California who have a doctoral degree which is available in the data, we can calculate the ratio by the following:

$$Ratio = \frac{\text{Total number of respondents}}{\text{Number of doctoral respondents}}$$

Once the ratio is known for California, it is assumed that this ratio is similar across other states. This is the core assumption of the ratio estimator: that the proportion of doctoral degree holders to total respondents is similar across states.

For each state, the estimated total number of respondents is calculated by applying the ratio derived from California:

$$\text{Estimated Total Respondents in State} = \frac{\text{Number of doctoral respondents in state}}{\text{Ratio}}$$

3 Results

3.1 Estimates and the actual number of respondents

Table 2 shows the number of estimated total respondents in each state by estimators approach of Laplace.

Table 2: Number of Estimated Total Respondents in Each State

State	Estimated Respondents
1	37042.7
2	10186.7
3	124340.0
4	15064.0
5	10927.6
6	8087.7
11	9384.2
12	88779.0
13	174656.4
14	100015.3
21	89952.0
22	38277.5
23	61182.2
24	74888.0
25	31671.5
31	15928.4
32	19817.8
33	35314.0
34	38339.2
35	9445.9
36	3704.3
37	4383.4
40	94520.6
41	28399.4
42	15496.2
43	168606.1
44	89581.6
45	27782.0
46	16237.1
47	87729.5
48	39944.4
49	198548.9
51	27658.6
52	99274.5
53	17348.3
54	51921.5
56	9816.3
61	55317.1
62	63651.7

Table 2: Number of Estimated Total Respondents in Each State

State	Estimated Respondents
63	10804.1
64	6976.4
65	17410.1
66	21608.2
67	26423.8
68	4445.1
71	391171.0
72	39944.4
73	73776.7
81	3148.6
82	13211.9
98	19200.5

Table 3 shows the actual respondent and the difference between estimation and the actual number of respondents in each state.

Table 3: Number of Actual Total Respondents and the Difference in Each State

State	Estimated Respondents	Actual Respondents	Difference
1	37042.7	37369	326.3
2	10186.7	14523	4336.3
3	124340.0	73077	-51263.0
4	15064.0	14077	-987.0
5	10927.6	10401	-526.6
6	8087.7	6860	-1227.7
11	9384.2	9641	256.8
12	88779.0	93166	4387.0
13	174656.4	203891	29234.6
14	100015.3	132605	32589.7
21	89952.0	128046	38094.0
22	38277.5	69843	31565.5
23	61182.2	101512	40329.8
24	74888.0	120666	45778.0
25	31671.5	61967	30295.5
31	15928.4	33586	17657.6
32	19817.8	29940	10122.2
33	35314.0	58984	23670.0
34	38339.2	64551	26211.8

Table 3: Number of Actual Total Respondents and the Difference in Each State

State	Estimated Respondents	Actual Respondents	Difference
35	9445.9	19989	10543.1
36	3704.3	8107	4402.7
37	4383.4	9296	4912.6
40	94520.6	88761	-5759.6
41	28399.4	51580	23180.6
42	15496.2	31288	15791.8
43	168606.1	217799	49192.9
44	89581.6	109349	19767.4
45	27782.0	45040	17258.0
46	16237.1	29796	13558.9
47	87729.5	109230	21500.5
48	39944.4	54651	14706.6
49	198548.9	292919	94370.1
51	27658.6	46605	18946.4
52	99274.5	62442	-36832.5
53	17348.3	39445	22096.7
54	51921.5	72374	20452.5
56	9816.3	18135	8318.7
61	55317.1	74153	18835.9
62	63651.7	59841	-3810.7
63	10804.1	19884	9079.9
64	6976.4	11116	4139.6
65	17410.1	30749	13338.9
66	21608.2	20243	-1365.2
67	26423.8	35537	9113.2
68	4445.1	5962	1516.9
71	391171.0	391171	0.0
72	39944.4	43708	3763.6
73	73776.7	80818	7041.3
81	3148.6	6972	3823.4
82	13211.9	14995	1783.1
98	19200.5	6718	-12482.5

4 Discussion

4.1 Reason of Difference

As the primary assumption is that the ratio of doctoral degree holders to total respondents in California is similar to that in other states. If this is not the case, the ratio estimator can produce biased estimates. Differences in educational attainment by state due to factors like demographics, local policies, and economy may cause the actual ratios to differ, leading to discrepancies between the estimates and actual values.

5 References