Estimating Statewide Respondent Totals Using the Laplace Ratio Estimator: A Case Study with 2022 ACS Data on Doctoral Degree Holders*

Arav Agarwal Daniel Du Dong Jun Yoon Grace Nguyen

November 21, 2024

Data

This repository uses data from IPUMS (*IPUMS USA* (2022)), and is analyzed with R (R Core Team (2023)). To obtain the data, follow these steps:

1. Google "IPUMS USA" and go to the first link: https://usa.ipums.org/usa/.
2. Click "Get Data" on the home screen.
3. Click "Select Samples". In USA Samples, tick the box that says "Default sample from each year", and then tick the box next to 2022 ACS. This should the only box ticked. Afterwards, click "submit sample selection".
4. Under Select Harmonized Variables, click the Household drop down menu and select Geographic. Click the + under "Add to cart" next to variable `STATEICP`.
5. Do the same for variable `SEX` in Demographic under the Person menu, and for variable `EDUC` in Education under the Person menu.
6. Click "View Cart". You should be taken to the extract request, with 1 sample and 14 variables. Change the data format to .csv, and then press "Submit Extract".
7. If you are not logged in to an account, you will be prompted to log in, or create an account. Once you are finished this, you can view and download the csv extract. There is also a link to the codebook that explains the variables in the dataset.

By using the codebook, we filter the amount of respondents in each state `STATEICP` that had a doctoral degree in `EDUCD` as their highest educational attainment. Here are the results in a table:

*Code and data are available at: IPUMS USA.

Table 1: Respondents In Each State with Doctoral Degree

State Code (STATEICP)	Number of Respondents
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
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

Overview of ratio estimators approach

Background

The purpose of the ratio estimator is to estimate a population parameter (such as a total or mean) using auxiliary information that is correlated with the variable of interest. By leveraging this auxiliary variable, the ratio estimator improves the precision of the estimate compared to simpler methods, like the sample mean. Rather than directly estimating the total number of respondents across all states, the ratio estimator uses a known ratio between two quantities—such as the number of people with doctoral degrees and the total number of respondents in California—and applies this ratio to other states. This method assumes that the relationship observed in California is approximately true for other states.

Approach

We are given the total number of respondents in California, 391171, as well as the total doctoral respondents in each state. To estimate the total number of respondents in each state we first calculate the following ratio:

Number of Respondents in a State : Number of Doctoral Respondents

That is, Ratio = 391171/n_Doctoral_CA

Then we extrapolate using this ratio by applying this ratio to our column of the number of respondents with a doctoral degree in each state. That is, we multiply each value in the collumn with the ratio we calculated.

Estimating Total Respondents

If, for example, there are 391,171 respondents in California (state 71 according the codebook) across all levels of education, we can use the ratio estimator to estimate the total number of respondents in each state as follows:

California respondents: 391171

California respondents with a doctorate: 6336

Ratio: 391171/6336 = 61.738

Table 2: Estimated Respondents vs Total Respondents in Each State

State Code (STATEICP)	Actual # of Respondents	Estimated # of Respondents
1	37043	37369
2	10187	14523
3	124340	73077
4	15064	14077
5	10928	10401
6	8088	6860
11	9384	9641
12	88779	93166
13	174656	203891
14	100015	132605
21	89952	128046
22	38277	69843
23	61182	101512
24	74888	120666
25	31672	61967
31	15928	33586
32	19818	29940
33	35314	58984
34	38339	64551
35	9446	19989
36	3704	8107
37	4383	9296
40	94521	88761
41	28399	51580
42	15496	31288
43	168606	217799
44	89582	109349
45	27782	45040
46	16237	29796
47	87729	109230
48	39944	54651
49	198549	292919
51	27659	46605
52	99274	62442
53	17348	39445
54	51922	72374
56	9816	18135
61	55317	74153
62	63652	59841
63	10804	19884
64	6976	11116
65	17410	30749
66	21608	20243
67	26424	35537
68	4445	5962
71	391171	391171
72	39944	43708
73	73777	80818
81	3149	6972
82	13212	14995
98	19200	6718

Any errors in the estimates are likely because of the following data limitations, discussed below:

Homogeneity Assumption:

The biggest assumption here is that the ratio of doctoral degree holders to total respondents is consistent across all states, which might not be true due to various socio-economic, educational, and cultural differences.

Variability in Smaller Populations:

States with very small populations or low numbers of doctoral degree holders might see more variability and less accuracy in the estimates.

References

IPUMS USA. 2022. 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/.