IPUMS 2022 ACS Analysis

Dhruv Gupta Justin Klip Kevin Shen

November 21, 2024

The IPUMS data service was used to extract data regarding college graduates by state, and doctoral degrees holders by state. We use the Laplace estimation technique to obtain an estimate doctoral degrees by state, given that there were 391,171 respondents in California. We compare our estimates for states we assume education levels are relatively high (Massachussets, New York), to states that are perceived to be uneducated (Arkansas, Alabama), and compare our estimates to actual data from IPUMS. We find that highly educated states, like Massuchusetts are overestimated, likely because of the high number of PhDs in their population relative to California. Further, we find that for uneducated states like Arkansas, we underestimate the number of respondents.

Instructions to Obtain the Data

In order to acquire the data used for our IPUMS research study, we first navigated to (https://www.ipums.org/). Once on the webpage, click on the "Visit Site" button under IPUMS USA. Next, click on the white "Get Data" button under CREATE YOUR CUSTOM DATA SET.

Once on the webpage used for selecting samples and variables, click on the blue "Select Samples" button on the left. On the USA Samples tab, uncheck the "Default sample from each year" checkbox, and then reselect the checkbox corresponding to 2022. Once selected, click the "Submit Samples" button; this will take you to the variable selection webpage.

On this page, navigate to the "Household" dropdown menu under Select Harmonized Variables, and select "Geographic." In the list of variables under Geographic Variables — Household, click the plus sign next to the "STATEICP" variable. Then, hover over the "Person" dropdown menu and select "Education." In the list of variables under Education Variables — Person, click the plus sign next to "EDUC."

In the Data Cart floating window, select "View Cart." On the Data Cart webpage, select the blue "Create Data Extract" button. On the Extract Request webpage, to the right of Data

Format, click "Change" and switch the format to .csv. Finally, press the "Submit Extract" button at the bottom of the page.

On the Download or Revise Extracts webpage, find the extract you requested, and in the Download Data column, click "Download .csv." This will give you the data necessary to run the code for our analysis.

Following the steps above, we obtained geographical data regarding college graduates by state, and doctoral degrees in that state from the IPUMS data service (IPUMS 2022). The analaysis for this data is done using (R Core Team 2023), and the packages used are Tidyverse, (Wickham et al. 2019) and KableExtra, (Zhu 2021).

Overview of Ratio Estimator Approach

We try to use the ratio estimators' approach to find total respondents for each state, given the fact that there are 391,171 respondents in California at all levels of education. We look at the number of total doctoral degree holders in California and compare it to the total number of respondents in California. We find there are about 61.74 times as many people compared to doctoral degrees in California. So, we simply multiply the number of doctoral degrees holders in each state by this ratio (61.74) to get our estimate of total respondents per state. The assumption is that other states would have similar ratios of PhD's to total respondents. This approach is necessary if we only have doctoral data for all the states, but not total respondent data.

Comparison of Estimated Totals and Actual Totals

Table 1 compares our estimates of doctoral degree holders with actual totals for relevant states. The full table, Table 2, be found in the appendix. We display Massachussets (StateICP = 3), New York (StateICP = 13), Alabama (StateICP = 41), Arkansas (StateICP = 42), Wyoming (StateICP = 68), and California (StateICP = 71).

Table 1: Sample Comparison of Estimated and Actual Total Respondents for Selected States

State ICP	Doctoral Count	Estimated Total Respondents	Actual Total Respondents
3	2014	124340.02	73077
13	2829	174656.37	203891
41	460	28399.41	51580
42	251	15496.20	31288
68	72	4445.12	5962
71	6336	391171.00	391171

Explanation as to Differing Results:

Some states may get differing values due to differing state characteristics. We based our ratio on California, which has a large number of educational institutes that award PhD's. California's large academic and advanced labor market may also draw in a lot of PhD's in comparison to a state like Wyoming, which has less incentive for PhD's to live there. This differing state level trends may explain why we get close estimates for some states, but not for others. In Wyoming (Code 68) we predict there to be a much smaller number of people (4445.125) than there actually are (5962) and this is because of this discrepancy in assuming there are much more PhD's in Wyoming then there are, so we multiply by too high of a number. Other states also underestimate suggesting the same issue as Wyoming. Massachusetts (Code 03), has an overestimate of the number of people, this is likely because they have a higher ratio of PhD graduates in their total population. This makes sense given that they have some of the best research institutions in the world that attract PhD's, so in Massachusetts our ratio is an underestimate.

Appendix

Table 2: Complete Comparison of Estimated and Actual Total Respondents by State

State ICP	Doctoral Count	Estimated Total Respondents	Actual Total Respondents
1	600	37042.71	37369
2	165	10186.74	14523
3	2014	124340.02	73077
4	244	15064.03	14077
5	177	10927.60	10401
6	131	8087.66	6860
11	152	9384.15	9641
12	1438	88779.02	93166
13	2829	174656.37	203891
14	1620	100015.31	132605
21	1457	89952.04	128046
22	620	38277.47	69843
23	991	61182.21	101512
24	1213	74888.01	120666
25	513	31671.52	61967
31	258	15928.36	33586
32	321	19817.85	29940
33	572	35314.05	58984
34	621	38339.20	64551
35	153	9445.89	19989
36	60	3704.27	8107
37	71	4383.39	9296
40	1531	94520.64	88761
41	460	28399.41	51580
42	251	15496.20	31288
43	2731	168606.06	217799
44	1451	89581.62	109349
45	450	27782.03	45040
46	263	16237.05	29796
47	1421	87729.48	109230
48	647	39944.39	54651
49	3216	198548.92	292919
51	448	27658.56	46605

52	1608	99274.46	62442
53	281	17348.34	39445
54	841	51921.53	72374
56	159	9816.32	18135
61	896	55317.11	74153
62	1031	63651.72	59841
63	175	10804.12	19884
64	113	6976.38	11116
65	282	17410.07	30749
66	350	21608.25	20243
67	428	26423.80	35537
68	72	4445.12	5962
71	6336	391171.00	391171
72	647	39944.39	43708
73	1195	73776.73	80818
81	51	3148.63	6972
82	214	13211.90	14995
98	311	19200.47	6718

Bibliography

IPUMS. 2022. "Integrated Public Use Microdata Series: USA, 2022 American Community Survey." Minnesota Population Center, University of Minnesota. https://www.ipums.org/.

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, et al. 2019. *Welcome to the Tidyverse*. https://doi.org/10.21105/joss.01686.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.