Analysis of ACS 2022

Lexun Yu Colin Sihan Yang Siddharth Gowda
Tanmay Shinde

November 21, 2024

This document provide instructions on downloading 2022 ACS data, a brief overview of the ratio estimators approach, estimates and the actual number of respondents and our explanation of why we think they are different.

1 Introduction

This paper is written with the help of R (R Core Team 2023). All the data are gathered from IPUMS USA (Ruggles et al. 2024). The code sections are done with the help of tidyverse (Wickham et al. 2019), janitor (Firke 2023) and knitr (Xie 2014).

2 Downloading Data

We followed the steps below to download the data:

1. Visit IPUMS USA at: https://usa.ipums.org/
2. Click on "Get Data" under "Create Your Custom Data Set"
3. Use the filter in "Select Harmonized Variables" and select "Geographic" under "Household". Check the "STATEICP" checkbox.
4. Use the filter in "Select Harmonized Variables" and select "Demographic" under "Person". Check the "SEX" checkbox.
5. Use the filter in "Select Harmonized Variables" and select "Education" under "Person". Check the "EDUC" checkbox.
6. Select "Select Samples", check only "ACS" in 2022, and click "Submit Sample Selections"

7. Click "View Cart" on the top of the page, click on "Create Data Extract". Change the data format to "Comma delimited (.csv)" and apply.

8. Submit extract, login and wait for the confirmation email.

3 Estimating the total numner of respondents

The ratio estimator method is a technique used to estimate the population total of interest in one state by leveraging the relationship between two related variables in a reference group.

In our case, we are given information about the number of respondents with doctoral degrees in each state, as shown in Table 1. We also know the total number of respondents in California (391,171). The ratio between the number of people with doctoral degrees and the total number of respondents in California can be used to estimate the total number of respondents in other state.

We follow these steps to estimate the total number of respondents in all states.

1. The ratio in California is calculated as the number of people with doctoral degrees in California divided by the total number of respondents in California: Ratio in California $= \frac{\text{Doctoral Count in California}}{\text{Total Respondents in California}} = \frac{6336}{391171} \approx 0.0161975$. The result is displayed in "Doctor Proportion" on Line 71 in Table 1.

2. We can then use this ratio to estimate the total number of respondents in other states. For any state x, the estimated total number of respondents is calculated by: Estimated Total Respondents in State $x = \frac{\text{Doctoral Count in State } x}{\text{Ratio in California}}$. The result is shown in Table 2.

The comparison between the estimated and actual data is displayed in Table 2.

4 Question 1

4.1 How many respondents were there in each state (stateicp) that had a doctoral degree as their highest educational attainment (EDUC)?

Table 1: Number of Respondents and Proportion of Doctors Per State

State ICP	Respondent Count	Doctor Count	Doctor Proportion
1	37369	600	0.016
2	14523	165	0.011
3	73077	2014	0.028
4	14077	244	0.017

Table 1: Number of Respondents and Proportion of Doctors Per State

State ICP	Respondent Count	Doctor Count	Doctor Proportion
5	10401	177	0.017
6	6860	131	0.019
11	9641	152	0.016
12	93166	1438	0.015
13	203891	2829	0.014
14	132605	1620	0.012
21	128046	1457	0.011
22	69843	620	0.009
23	101512	991	0.010
24	120666	1213	0.010
25	61967	513	0.008
31	33586	258	0.008
32	29940	321	0.011
33	58984	572	0.010
34	64551	621	0.010
35	19989	153	0.008
36	8107	60	0.007
37	9296	71	0.008
40	88761	1531	0.017
41	51580	460	0.009
42	31288	251	0.008
43	217799	2731	0.013
44	109349	1451	0.013
45	45040	450	0.010
46	29796	263	0.009
47	109230	1421	0.013
48	54651	647	0.012
49	292919	3216	0.011
51	46605	448	0.010
52	62442	1608	0.026
53	39445	281	0.007
54	72374	841	0.012
56	18135	159	0.009
61	74153	896	0.012
62	59841	1031	0.017
63	19884	175	0.009
64	11116	113	0.010
65	30749	282	0.009
66	20243	350	0.017

Table 1: Number of Respondents and Proportion of Doctors Per State

State ICP	Respondent Count	Doctor Count	Doctor Proportion
67	35537	428	0.012
68	5962	72	0.012
71	391171	6336	0.016
72	43708	647	0.015
73	80818	1195	0.015
81	6972	51	0.007
82	14995	214	0.014
98	6718	311	0.046

4.2 Laplace to estimator. Given 391,171 California Respondents use estimator to find number of participants for all states.

4.2.1 Different between estimator and non estimator

Table 2: Esimated Number of Respondents Per State

State ICP	Number of Doctor	Estimated Respondent Count	Respondent Count	Difference
1	600	37043	37369	-326
2	165	10187	14523	-4336
3	2014	124340	73077	51263
4	244	15064	14077	987
5	177	10928	10401	527
6	131	8088	6860	1228
11	152	9384	9641	-257
12	1438	88779	93166	-4387
13	2829	174656	203891	-29235
14	1620	100015	132605	-32590
21	1457	89952	128046	-38094
22	620	38277	69843	-31566
23	991	61182	101512	-40330
24	1213	74888	120666	-45778
25	513	31672	61967	-30295
31	258	15928	33586	-17658
32	321	19818	29940	-10122
33	572	35314	58984	-23670
34	621	38339	64551	-26212
35	153	9446	19989	-10543

Table 2: Esimated Number of Respondents Per State

State ICP	Number of Doctor	Estimated Respondent Count	Respondent Count	Difference
36	60	3704	8107	-4403
37	71	4383	9296	-4913
40	1531	94521	88761	5760
41	460	28399	51580	-23181
42	251	15496	31288	-15792
43	2731	168606	217799	-49193
44	1451	89582	109349	-19767
45	450	27782	45040	-17258
46	263	16237	29796	-13559
47	1421	87729	109230	-21501
48	647	39944	54651	-14707
49	3216	198549	292919	-94370
51	448	27659	46605	-18946
52	1608	99274	62442	36832
53	281	17348	39445	-22097
54	841	51922	72374	-20452
56	159	9816	18135	-8319
61	896	55317	74153	-18836
62	1031	63652	59841	3811
63	175	10804	19884	-9080
64	113	6976	11116	-4140
65	282	17410	30749	-13339
66	350	21608	20243	1365
67	428	26424	35537	-9113
68	72	4445	5962	-1517
71	6336	391171	391171	0
72	647	39944	43708	-3764
73	1195	73777	80818	-7041
81	51	3149	6972	-3823
82	214	13212	14995	-1783
98	311	19200	6718	12482

Based on the Table 2, the estimator is usually less than the actual. This is probably the case because California is more educated compared to most American states. Thus, they will have a higher percentage of doctors, resulting in an underestimate of the total number of respondents.

References

Firke, Sam. 2023. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

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

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. 2024. "IPUMS USA: Version 15.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.

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.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.