2022 ACS Analysis using Ratio Estimators

Mandy He, Lorina Yang, Ruiying Li, Wendy Yuan, Benjamin Fleurence

October 3, 2024

Introduction

The following packages in R (R Core Team (2023)) were used in the data analysis process: tidyverse (Wickham et al. (2023b)), dplyr (Wickham, François, et al. (2023)), knitr (Xie (2021)), readr (Wickham, Hester, and Bryan (2024)), and ggplot2 (Wickham et al. (2023a)).

Instructions of how to obtain the data

1. Create a user/log into IPUMS account
2. Navigate to IPUMS USA site
3. Create a custom data set by clicking on the "Get Data" button
4. Select samples for only ACS 2022 and submit sample selection from (Ruggles et al. 2021)
5. Under select harmonized variables select 'geographic' from the 'household' dropdown menu
    1. Select the 'STATEICP' variable
6. Under select harmonized variables select 'demographic' from the 'person' dropdown menu
    1. Select the 'SEX' variable
7. Under select harmonized variables select 'education' from the 'person' dropdown menu
    1. Section the 'EDUC' variable
8. Click view cart and create data extract
9. Set up data format to be .csv and submit extract
10. Wait for data to be processed for download
11. Download the .csv when the status of the data is completed
12. Read the datafile once downloaded using the read_csv() function in R from the readr package

Ratio Estimators Approach

Ratio estimator is a statistical method used to estimate a total population by looking at the ratio between two related variables in a sample, since it is difficult to measure the entire population directly. By finding the ratio of a specific characteristic to the total sample size, this ratio can then be applied to estimate the full population based on available sample data.

In our case, we aim to estimate the total number of respondents in each state using the ratio estimator approach, based on the number of individuals with doctoral degrees (EDUCD) in California. Since we know that California has a total of 391,171 respondents across all education levels and the number of respondents with doctoral degrees, we can calculate the ratio of doctoral degree holders to the total population in California. Assuming this ratio is similar across other states, we can then use it to estimate the total number of respondents in each state. This is done by dividing the number of doctoral degree holders in each state by the ratio calculated for California, giving an estimate of the total population of respondents for each state.

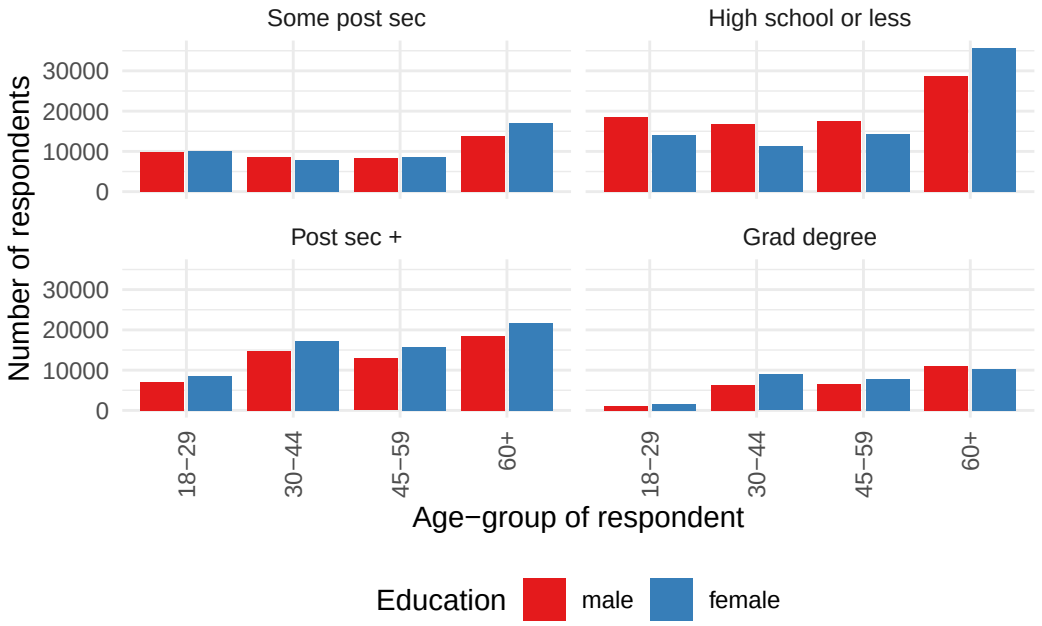

Figure 1: Number of Respondents Based on Education Level and Age

In Table 1 ,the estimate is significantly greatly then the actual count of number of respondents with doctoral degrees. This is likely because we assumed that the ratio for California is similar to the ratio for other states.

Table 1: Difference Estimate vs Actual Number of Respondants per State

stateicp	doctoral_count	estimated_total	actual_count	difference
connecticut	88	37014	4610	32404
maine	19	7992	1860	6132
massachusetts	303	127446	9056	118390
new hampshire	38	15983	1774	14209
rhode island	32	13460	1321	12139
vermont	25	10515	884	9631
delaware	20	8412	1184	7228
new jersey	215	90432	11263	79169
new york	385	161936	25282	136654
pennsylvania	227	95479	16317	79162
illinois	214	90011	15548	74463
indiana	95	39958	8403	31555
michigan	151	63513	12418	51095
ohio	199	83702	14646	69056
wisconsin	86	36173	7571	28602
iowa	45	18928	4028	14900
kansas	46	19348	3496	15852
minnesota	73	30705	7066	23639
missouri	87	36593	7795	28798
nebraska	27	11357	2381	8976
north dakota	13	5468	962	4506
south dakota	13	5468	1074	4394
virginia	232	97582	10842	86740
alabama	73	30705	6404	24301
arkansas	41	17245	3708	13537
florida	434	182546	27447	155099
georgia	210	88329	13118	75211
louisiana	68	28602	5474	23128
mississippi	32	13460	3583	9877
north carolina	206	86646	13370	73276
south carolina	103	43323	6704	36619
texas	487	204839	34990	169849
kentucky	68	28602	5581	23021
maryland	241	101368	7503	93865
oklahoma	46	19348	4622	14726

Table 1: Difference Estimate vs Actual Number of Respondants per State

stateicp	doctoral_count	estimated_total	actual_count	difference
tennessee	119	50053	8829	41224
west virginia	18	7571	2229	5342
arizona	137	57624	9004	48620
colorado	160	67298	7299	59999
idaho	32	13460	2301	11159
montana	16	6730	1316	5414
nevada	46	19348	3669	15679
new mexico	55	23134	2456	20678
utah	66	27761	3886	23875
wyoming	13	5468	718	4750
california	930	391171	47881	343290
oregon	93	39117	5410	33707
washington	190	79917	9940	69977
alaska	6	2524	786	1738
hawaii	35	14721	1838	12883
district of columbia	55	23134	899	22235

Difference in estimate vs actual number of respondents

1. Sampling variability

The ratio estimator is based on the assumption that the ratio of doctoral degree holders to the total respondents in California applies to all other states. However, the true ratio may differ from state to state due to differences in educational attainment across regions. Some states might have a higher or lower proportion of doctoral degree holders compared to California, leading to inaccuracies when applying the California ratio to other states.

2. Non-uniform education attainment

Educational attainment, including the proportion of respondents with doctoral degrees, is not uniform across the U.S. Factors such as state demographics, local economies, and educational infrastructure affect the number of people with higher degrees. States with large research universities (e.g., Massachusetts) may have more doctoral degree holders, whereas others may have fewer.

3. Survey non-response

Not all selected participants respond to the ACS survey. If the non-response rate is high and if those who don't respond differ systematically from those who do (for instance, in their level of education), the actual total counts could deviate from the estimates. The ratio estimator does not account for non-response bias.

Reference

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, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. 2021. "IPUMS USA: Version 11.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/d010.v11.0.

Wickham, Hadley et al. 2023a. *Ggplot2: Elegant Graphics for Data Analysis*. https://CRAN.R-project.org/package=ggplot2.

——— et al. 2023b. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.