Summary of Respondents and Doctoral Counts with Estimation by State for 2022 ACS*

Yiyi Feng Sakura Hu

November 21, 2024

This paper examines state-level population estimates using the 2022 American Community Survey (ACS) data from IPUMS USA. The study applies the ratio estimator method, using the ratio between the number of respondents and doctoral degree in California as a baseline, to estimate the total number of respondents across all states. Comparing these estimates with actual respondent counts revealed significant discrepancies, largely due to variations in population demographics and sampling distributions across states. These findings have shown the limitations of ratio-based estimations in contexts with diverse demographic and sampling characteristics.

Table of contents

*Code and data are available at: https://github.com/kqlqkqlqF/Summary-of-Respondents-for-2022-ACS.git.

1 Introduction

Accurately estimating state-level populations using sample data is a important task in demographic research, with implications for policy making, resource allocation, and regional planning. The 2022 American Community Survey (ACS) dataset (Ruggles et al. 2024), accessed via IPUMS USA, provides detailed information on individual respondents, enabling state-by-state analyses of population characteristics. However, estimating total populations across states using ratio-based methods introduces challenges due to varying demographic and educational dynamics across regions.

This paper applies the ratio estimator approach to estimate the total number of respondents in each U.S. state, using California's data as a baseline. The ratio of doctoral degree holders to total respondents in California was used to project total respondent counts for other states. These estimates were then compared to actual respondent numbers, revealing notable differences. These differences come from variations in state education policies, the distribution of educational resources, population age structures, and the influence of regional industries. For instance, states with a concentration of research institutions or industries requiring advanced education often exhibit higher proportions of doctoral degree holders, while demographic factors such as age and migration patterns further influence these variations.

The findings have shown the limitations of applying uniform ratio estimations across diverse populations and highlight the importance of accounting for regional demographic and economic differences in population estimation. Future studies could refine estimation techniques to better account for such heterogeneity, improving the accuracy and reliability of these methods.

The remainder of this paper is structured as follows: Section 2 outlines the process for obtaining and preparing the dataset. Section 3 describes the estimation methodology and the resulting comparisons with actual respondent data. Section 4 explains the observed differences and their underlying causes.

2 Data

We obtained the data from IPUMS USA (Ruggles et al. 2024) with the following steps: First, open the IPUMS website and navigate to the IPUMS USA section. Once there, create a data set by selecting samples from the '2022 ACS' dataset and submit the selection. Next, choose the following harmonized variables for your dataset: (i) STATEICP under GEOGRAPHIC VARIABLES–HOUSEHOLD, (ii) SEX under DEMOGRAPHIC VARIABLES–PERSON, and (iii) EDUC under EDUCATION VARIABLES–PERSON.

After selecting the variables, proceed to view your cart and create the data extract. Change the data format from the default .dat format to .csv format, then submit the extract. Wait until the status of your data extract changes to 'completed,' at which point you can download the data. Upon downloading, a file with the .gz suffix will be received. Decompress this file to obtain the final dataset in .csv format.

The packages used in this paper provide essential tools for data analysis and reporting. Package tidyverse (Wickham et al. 2019) offers a suite for data manipulation, visualization, and programming. Package dplyr (Wickham et al. 2023) simplifies data manipulation with tools for summarizing. Package knitr (Xie 2022) supports dynamic report generation, integrating code and results, while package readr (Wickham, François, and Henry 2023) imports rectangular text data. Finally, package testthat (Wickham 2023) allows for testing data processing steps to ensure accuracy.

Table 1: Summary of Respondents and Doctoral Counts with Estimation by State for 2022 ACS

State	Total Respondents	Total Actual Doctor	Estimated Respondents	Difference in Respondents
Alabama	51580	460	28399	23181
Alaska	6972	51	3149	3823
Arizona	74153	896	55317	18836
Arkansas	31288	251	15496	15792
California	391171	6336	391171	0
Colorado	59841	1031	63652	-3811
Delaware	9641	152	9384	257
District of Columbia	6718	311	19200	-12482
Florida	217799	2731	168606	49193
Georgia	109349	1451	89582	19767
Hawaii	14995	214	13212	1783
Idaho	19884	175	10804	9080
Illinois	128046	1457	89952	38094
Indiana	69843	620	38277	31566
Iowa	33586	258	15928	17658
Kansas	29940	321	19818	10122
Kentucky	46605	448	27659	18946
Louisiana	45040	450	27782	17258
Maryland	62442	1608	99274	-36832
Michigan	101512	991	61182	40330
Minnesota	58984	572	35314	23670
Mississippi	29796	263	16237	13559
Missouri	64551	621	38339	26212
Montana	11116	113	6976	4140
Nebraska	19989	153	9446	10543
Nevada	30749	282	17410	13339

State	Total Respondents	Total Actual Doctor	Estimated Respondents	Difference in Respondents
New Jersey	93166	1438	88779	4387
New Mexico	20243	350	21608	-1365
New York	203891	2829	174656	29235
North Carolina	109230	1421	87729	21501
North Dakota	8107	60	3704	4403
Ohio	120666	1213	74888	45778
Oklahoma	39445	281	17348	22097
Oregon	43708	647	39944	3764
Pennsylvania	132605	1620	100015	32590
South Carolina	54651	647	39944	14707
South Dakota	9296	71	4383	4913
Tennessee	72374	841	51922	20452
Texas	292919	3216	198549	94370
Utah	35537	428	26424	9113
Virginia	88761	1531	94521	-5760
Washington	80818	1195	73777	7041
West Virginia	18135	159	9816	8319
Wisconsin	61967	513	31672	30295
Wyoming	5962	72	4445	1517

3 Result

3.1 Estimation

We start by matching STATEICP to the state name getting the actual value for each state and replacing NA with 0. Select California's row from the actual values we get, using the number of doctoral degrees in the California data as a percentage of total respondents to get a ratio. Finally, the doctoral degree and the proportion of total respondents obtained in California are mapped to each state, and the estimated total respondents of each state are obtained from the doctoral degree of each state. The difference in respondents column is obtained by comparing the value we obtained with the actual value.

3.2 Summary of the Result

From Table 1, we can see significant differences in the estimated total number of respondents across U.S. states, based on the proportion of respondents with a PhD in California. In many states, the estimated difference is nearly half of the state's actual respondent count, such as in Alabama, Alaska, and Wisconsin. The estimates generally exceed the actual respondent count, except in Colorado, Washington D.C., New Mexico, Maryland, and Virginia, where the estimates are lower than the actual numbers.

4 Discussion

4.1 Explanation of the Difference

There are several reasons why the ratio of people holding a doctoral degree in each state can vary, such as state education policies or cultural attitudes toward education that lead to differences in the distribution of educational resources across states. For instance, states with more universities and research institutions tend to have a higher proportion of residents with advanced degrees. Therefore, using California's doctorate-to-respondent ratio as representative of all states may not be accurate, as it does not account for these regional differences in education opportunities and demographics.

Additionally, population demographics vary between states. States with more young individuals may have fewer people with advanced degrees, while states with older populations may have more. States with industries that require highly educated workers often have more individuals with advanced degrees, and states with a higher cost of living may also attract more individuals with advanced degrees due to higher earning potential.

References

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 [Dataset]." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.

Wickham, Hadley. 2023. *Testthat: Get Started with Testing in r*. https://CRAN.R-project.org/package=testthat.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. *Welcome to the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Romain François, and Lionel Henry. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.