US Doctorate Estimates

October 3, 2024

1 Introduction

In this paper we explore the US 2022 Census Data, sourced from Ruggles et al. (2024) We use R Core Team (2023), Wickham, Hester, and Bryan (2023), Wickham et al. (2023), Xie (2023)

2 Running Code

The data is obtained from IPSUMS USA. Once on their website, navigate to Get Data. This brings up a search function. First, click select samples, deselect the "Default sample from each year" tick, and manually tick 2022 ACS. This is data taken in the US, not quite at the scale of a census, but involves much more thorough questioning. Click on 'Submit Summary' to add this dataset. Under Harmonized Variables, select the following through the dropdown menus: Household -> State -> STATEICP, Person -> Demographic -> SEX, Person -> Education -> EDUC. Finally, click on 'View Cart' and follow the steps required to download data (an IPSUMS account is required when checking out).

```
 [1] "YEAR"     "SAMPLE"   "SERIAL"   "CBSERIAL" "HHWT"     "CLUSTER"
 [7] "STATEICP" "STRATA"   "GQ"       "PERNUM"   "PERWT"    "SEX"
[13] "EDUC"     "EDUCD"
```

The ratio estimator approach involves the ratio of two random variables (R = a_x/a_y. It is used to estimate the population given a ratio and a preexisting population value (a_y = R*a_x). In this case, we are using California's ratio of doctorates (R), and the number of correspondents in California (a_x) to estimate the number of correspondents in other states (a_y).

Table 1: Comparison of Actual and Estimated Total Respondents in Each State Based on Doctoral Degree Ratio

STATEICP	actual_total_respondents	doctoral_count	estimated_total_respondents
1	37369	600	37042.708
2	14523	165	10186.745
3	73077	2014	124340.024
4	14077	244	15064.035
5	10401	177	10927.599
6	6860	131	8087.658
11	9641	152	9384.153
12	93166	1438	88779.024
13	203891	2829	174656.370
14	132605	1620	100015.312
21	128046	1457	89952.043
22	69843	620	38277.465
23	101512	991	61182.207
24	120666	1213	74888.009
25	61967	513	31671.516
31	33586	258	15928.365
32	29940	321	19817.849
33	58984	572	35314.049
34	64551	621	38339.203
35	19989	153	9445.891
36	8107	60	3704.271
37	9296	71	4383.387
40	88761	1531	94520.644
41	51580	460	28399.410
42	31288	251	15496.200
43	217799	2731	168606.061
44	109349	1451	89581.616
45	45040	450	27782.031
46	29796	263	16237.054
47	109230	1421	87729.481
48	54651	647	39944.387
49	292919	3216	198548.917
51	46605	448	27658.556
52	62442	1608	99274.458
53	39445	281	17348.335
54	72374	841	51921.530
56	18135	159	9816.318
61	74153	896	55317.111

STATEICP	actual_total_respondents	doctoral_count	estimated_total_respondents
62	59841	1031	63651.720
63	19884	175	10804.123
64	11116	113	6976.377
65	30749	282	17410.073
66	20243	350	21608.247
67	35537	428	26423.799
68	5962	72	4445.125
71	391171	6336	391171.000
72	43708	647	39944.387
73	80818	1195	73776.727
81	6972	51	3148.630
82	14995	214	13211.899
98	6718	311	19200.470

Our estimated number of respondents were much higher than the actual number of respondents for most states that are not California. One reason could be that California's count of doctorates is much larger than most other states, which would inflate the correspondent estimator. If we had used another state for our estimator, one closer to the average number across all states, the estimator would get closer to the actual number.

References

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 Rogers, and Megan Schouweiler. 2024. "IPUMS USA: Version 15.0 [Dataset]." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.

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

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.