Reflection Exercise 4

Data was obtained through Integrated Public Use Microdata Series USA (IPUMS USA) (Ruggles et al. 2021) and contains data surrounding the 2022 American Community Survey (ACS). Two variables were extracted from the 2022 ACS: STATEICP and EDUC. After selecting the correct data and converting the file from a .dat to a .dta, you could download the data extract from IPUMS USA. You then unzip the file and import it into R (R Core Team 2023).

This analysis uses the dplyr, tidyr, and haven packages in order to synthesize the data (Wickham et al. 2019).

The ratio estimators approach was used in order to estimate the total number of respondents in each state. This approach allows you to take the ratio between two means for a subset of your data in which you know both values, and broad apply it to the broader pool in which you only have one of the two values needed. Given the actual number of respondents in California and after extracting the number of doctoral respondents in California, we were able to determine the ratio of respondents to doctoral respondents. This ratio was then applied to the number of doctoral respondent in every state, giving us an estimated total number of respondents per state.

In Table 1, we see the number of doctoral respondents, the estimated number of total respondents, and the actual number of total respondents per state.

In some instances the estimated amount of total respondents is very close to the actual value. In others, such as Massachusetts (STATEICP = 2) or Indiana (STATEICP = 22), we see large discrepancies. These numbers are so off in these states because the proportion of state populations who have doctoral degrees varies widely from state to state. In Massachusetts our estimated total is much higher than the actual because a larger proportion of respondents in Massachusetts have doctoral degrees compared to California. The same applies to Indiana but in the reverse, in which they have a lower proportion of doctoral respondents than California.

Table 1: The number of doctoral respondents vs estimated respondents vs actual respondents in each state

```
# A tibble: 51 x 4
   stateicp doctoral_count estimated_total actual_total
   <fct>             <int>           <dbl>        <int>
 1 1                   600           37043        37369
 2 2                   165           10187        14523
 3 3                  2014          124340        73077
 4 4                   244           15064        14077
 5 5                   177           10928        10401
 6 6                   131            8088         6860
 7 11                  152            9384         9641
 8 12                 1438           88779        93166
 9 13                 2829          174656       203891
10 14                 1620          100015       132605
11 21                 1457           89952       128046
12 22                  620           38277        69843
13 23                  991           61182       101512
14 24                 1213           74888       120666
15 25                  513           31672        61967
16 31                  258           15928        33586
17 32                  321           19818        29940
18 33                  572           35314        58984
19 34                  621           38339        64551
20 35                  153            9446        19989
21 36                   60            3704         8107
22 37                   71            4383         9296
23 40                 1531           94521        88761
24 41                  460           28399        51580
25 42                  251           15496        31288
26 43                 2731          168606       217799
27 44                 1451           89582       109349
28 45                  450           27782        45040
29 46                  263           16237        29796
30 47                 1421           87729       109230
31 48                  647           39944        54651
32 49                 3216          198549       292919
33 51                  448           27659        46605
34 52                 1608           99274        62442
35 53                  281           17348        39445
36 54                  841           51922        72374
37 56                  159            9816        18135
38 61                  896           55317        74153
39 62                 1031           63652        59841
40 63                  175           10804        19884
41 64                  113            6976        11116
42 65                  282           17410        30749
43 66                  350           21608        20243
44 67                  428           26424        35537
45 68                   72            4445         5962
46 71                 6336          391171       391171
47 72                  647           39944        43708
```

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, 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, 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.