Reflection Exercise 4

Ying Zhang

Please use IPUMS to access the 2022 ACS.

1. Instructions on how to obtain the data.

To obtain the data from IPUMS USA, go to IPUMS, then “IPUMS USA”, and click “Get Data”. We are interested in a sample, so go to “SELECT SAMPLE”. Un-select “Default sample from each year” and instead select “2022ACS” and then “SUBMIT SAMPLE SELECTIONS”. We might be interested in data based on state. We would begin by looking at “HOUSEHOLD” variables and selecting “GEOGRAPHIC”. We add “STATEICP” to our “cart” by clicking the plus, which will then turn into a tick. We might then be interested in data on a “PERSON” basis, for instance, “DEMOGRAPHIC” variables such as “AGE”, which we should add to our cart. We also want “SEX” and “EDUC” (both are in “PERSON”). When we are done, we can click “VIEW CART”, and then click “CREATE DATA EXTRACT”. At this point there are two aspects that we likely want to change: 1. change the “DATA FORMAT” from “.dat” to “.csv”; 2. Customize the sample size as we likely do not need three million responses, and could just change it to, say, 500,000. Briefly check the dimensions of the request. It should not be much more than around 40MB. If it is then check whether there are variables accidentally selected that are not needed or further reduce the number of observations. Finally, we want to include a descriptive name for the extract, for instance, “2024-10-03: State, age, sex, education”, which specifies the date we made the extract and what is in the extract. After that we can click “SUBMIT EXTRACT”.

Making use of the codebook, how many respondents were there in each state (STATEICP) that had a doctoral degree as their highest educational attainment (EDUC)? (Hint: Make this a column in a tibble.)

```r
#           101
doctorate_code <- 116

#
respondents_by_state <- ipums_extract %>%
  filter(EDUCD == doctorate_code) %>%  #
  group_by(STATEICP) %>%                  #
  summarise(count = n()) %>%             #
  as_tibble()                             #    tibble

#
print(respondents_by_state)
```

```
# A tibble: 51 x 2
   STATEICP count
      <dbl> <int>
 1        1   600
 2        2   165
 3        3  2014
 4        4   244
 5        5   177
 6        6   131
 7       11   152
 8       12  1438
 9       13  2829
10       14  1620
# i 41 more rows
```

If I tell you that there were 391,171 respondents in California (STATEICP) across all levels of education, then can you please use the ratio estimators approach of Laplace to estimate the total number of respondents in each state i.e. take the ratio that you worked out for California and apply it to the rest of the states. (Hint: You can now work out the ratio between the number of respondents with doctoral degrees in a state and number of respondents in a state and then apply that ratio to your column of the number of respondents with a doctoral degree in each state.)

2. A brief overview of the ratio estimators approach.

The ratio estimators approach is a statistical method used to estimate the total population or means based on known ratios derived from a sample. This technique involves calculating the ratio of a specific characteristic (such as the number of individuals holding a doctoral degree) to the total population within a known subset (such as California). This calculated ratio is then applied to other subsets to estimate their totals, under the assumption that similar relationships hold across the entire population. This approach is particularly beneficial when the exact population size is not available, but the sample provides proportional relationships that can be generalized.

Then compare it to the actual number of respondents in each state.

3. Your estimates and the actual number of respondents.

```
# A tibble: 51 x 3
   STATEICP actual_total estimated_total
      <dbl>        <int>           <dbl>
 1        1        37369           37043
 2        2        14523           10187
 3        3        73077          124340
 4        4        14077           15064
 5        5        10401           10928
 6        6         6860            8088
 7       11         9641            9384
 8       12        93166           88779
 9       13       203891          174656
```

```
10       14       132605          100015
# i 41 more rows
```

4. Some explanation of why you think they are different.

The estimated total number of respondents in each state using the ratio estimator may differ from the actual number of respondents primarily because this method assumes that the proportion of doctoral degree holders in California is applicable to other states. However, variations in educational attainment are influenced by various factors, including demographic characteristics, economic opportunities, and educational infrastructure, which can differ significantly across states. Furthermore, if the data used is a sample rather than a complete census, "sampling variability" will also affect the accuracy of the calculated ratio and its estimates. Additionally, the distribution of educational attainment in the U.S. is not uniform; factors such as regional policies, cultural differences, and access to higher education contribute to this "non-uniform distribution," meaning California's ratio may not be representative of other states. Lastly, the effectiveness of the Laplace ratio approach relies on the consistency of the relationship between the characteristic of interest and the population. If the ratio of doctoral degree holders to the total population in California does not hold for other states due to unobserved factors, this could introduce "bias in the ratio." These considerations highlight that the "assumption of similarity" inherent in ratio estimators often leads to discrepancies when applied to diverse populations, such as different states.