reflection 10.3.24

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)?

```
# A tibble: 51 x 2
   STATEICP doctoral_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
```

Instructions on obtaining the data.

To collect data from IPUMS USA, we first navigated to the IPUMS website and selected “IPUMS USA.” We then clicked on “Get Data” and chose the “2022 ACS” sample under the “SELECT SAMPLE” section. To gather state-level data, we selected “HOUSEHOLD” followed by “GEOGRAPHIC” and added “STATEICP” to the cart. For individual-level data, we went to the “PERSON” section and added “EDUC” to the cart. After reviewing our selections by clicking “VIEW CART,” we proceeded to “CREATE DATA EXTRACT.” We changed the “DATA FORMAT” to “.csv” and clicked “SUBMIT EXTRACT.” After logging in with the account, we received an email when the extract was ready for download. Finally, we downloaded the file to use in RStudio.

A brief overview of the ratio estimators approach.

The ratio estimator is a method used in survey sampling to improve estimation accuracy by leveraging a known relationship between two variables. This method calculates the ratio of a particular attribute to the total population for a known subset. The ratio is then applied to other subsets to approximate totals, assuming similar correlations exist across the entire population. It is especially helpful when the precise population size is unknown but a sample yields proportional connections.

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
```

Explanation of why they are different.

The estimated total number of respondents in each state using the ratio estimator can differ from the actual count due to several factors:

1. Assumption of Similarity: The ratio estimator makes the assumption that the percentage of Californians with doctorates is typical of other states; however, due to a variety of reasons, including economic, demographic, and educational infrastructure, there are considerable variations in educational attainment.
2. Sampling Variability: If based on a sample, random variability can impact the ratio and estimation accuracy.
3. Non-Uniform Distribution: Educational attainment isn't evenly distributed across the U.S., so any areas' ratio may not apply to other states.

4. Bias: When relationships are constant over all domains, the ratio technique performs well. The estimations will be skewed if the ratio is impacted by unobserved factors.

These factors explain why using the ratio estimator across diverse states often leads to differences from actual numbers.