Estim Counts of Education Level: A 2022 ACS Analysis

```
library(haven)
library(labelled)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
ipums_extract <- read_csv("usa_00004.csv.gz")
```

```
Rows: 3373378 Columns: 13
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (13): YEAR, SAMPLE, SERIAL, CBSERIAL, HHWT, CLUSTER, STATEICP, STRATA, G...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ipums_extract <-
  ipums_extract |>
  select(STATEICP, EDUCD) |>
```

```
  rename(state_code = STATEICP,
         education_code = EDUCD) |>
  to_factor()
```

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.)

```
phd_data <- ipums_extract |>
  filter(education_code == 116) |>
  group_by(state_code) |>
  summarise(phd_count = n()) |>
  ungroup()
phd_data
```

```
# A tibble: 51 x 2
   state_code phd_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 how to obtain the data.

To the 2022 ACS data from IPUMS USA, visit the IPUMS USA website and select "IPUMS USA." Click "Get Data," choose "2022 ACS," and add "STATEICP" under "HOUSEHOLD" > "GEOGRAPHIC" and "EDUCD" under "PERSON" to your cart. View your cart and create a data extract, selecting your desired format like ".dta". Submit the extract, log in or create an account, and download the file once ready. Save it locally for analysis.

A brief overview of the ratio estimators approach.

Ratio are a statistical method to estimate population totals by leveraging known sample ratios. The technique calculates the ratio of a characteristic, like doctoral degree holders, to the total population in a subset, such as California, and applies this ratio to other subsets, assuming consistent proportional relationships. Useful in fields like economics and demography, it allows for informed estimates without full population data. The method hinges on accurate assumptions that the initial sample reflects broader patterns. Despite challenges, ratio estimators are cost-effective for understanding large populations, supporting policy planning and resource allocation with proper validation.

Your estimates and the actual number of respondents.

```
# Total number of participants from California (given value)
total_participants_ca <- 391171

# Get the count of participants with a PhD in California
phd_participants_ca <- phd_data |>
  filter(state_code == 71) |>
  pull(phd_count)

# Calculate the proportion of PhD holders relative to the total number of participants in Ca
phd_proportion_ca <- phd_participants_ca / total_participants_ca

# Estimate the total number of participants in each state using the proportion estimator
estimated_totals <- phd_data |>
  mutate(predicted_total = phd_count / phd_proportion_ca)

# Aggregate the actual count of participants by state
actual_totals <- ipums_extract |>
  group_by(state_code) |>
  summarise(actual_total = n()) |>  # Count of actual participants by state
  ungroup()

# Merge the predicted and actual participant counts for comparison
totals_comparison <- phd_data |>
  left_join(actual_totals, by = "state_code") |>
  left_join(estimated_totals, by = "state_code") |>
  select(state_code, actual_total, predicted_total)
```

```
# Display the comparison between actual and predicted participant counts
totals_comparison
```

```
# A tibble: 51 x 3
   state_code actual_total predicted_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
```

Some explanation of why you think they are different.

The ratio estimators approach may differ from actual numbers due to several reasons:

- Assumption of Uniformity:

  The method presumes that the ratio of doctoral degree holders to total respondents in California accurately reflects conditions in other states. However, educational attainment can vary significantly due to state policies, access to higher education, and socio-economic factors, leading to misestimated totals.

- Demographic Variability:

  Each state has distinct demographic characteristics, including age distribution, income levels, cultural priorities, and historical developments influencing education. These differences mean that educational patterns observed in one state may not apply to others, affecting the accuracy of ratio-based estimates.

- Sample Size and Composition:

  The initial sample from California may not be sufficiently large or representative. If the sample doesn't capture the diversity of the population, extrapolating ratios from it could result in inaccurate estimates for states with different demographic compositions.

- Local Factors:

  States may have unique factors affecting the presence of doctoral degree holders, such as the density of universities and colleges, employment opportunities requiring advanced degrees, and migration trends of highly educated individuals.

- Policy and Economic Influences:

  Differences in state-level education funding, scholarship availability, and job markets can greatly influence the number of individuals pursuing and obtaining doctoral degrees.

These nuances underscore the importance of careful validation and adjustment of assumptions when using ratio estimators. Utilizing additional data sources and considering local contexts can help refine estimates and enhance their reliability.