STA 322 Project 1

Sam Hamelsky, Carli Olsen, Anmol Sapru

```
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   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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 becom
```

```
library(survey)
```

```
Loading required package: grid
Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack

Loading required package: survival

Attaching package: 'survey'

The following object is masked from 'package:graphics':

    dotchart
```

```r
library(sampling)
```

```
Attaching package: 'sampling'

The following objects are masked from 'package:survival':

    cluster, strata
```

The data was taking directly from the scorecard website. To make the code more reproducible and without issues of downloading and putting the data in the correct place, the data is saved straight from online. There is both the college scorecard, which gives a single row to each college and gives meta-information about each school and then the field of study, which has a row per each of the department per college.

```r
url <- "https://ed-public-download.scorecard.network/downloads/Most-Recent-Cohorts-Institutio

college_scoreboard <- tempfile(fileext = ".zip") |>
  {\(tmp) {
    download.file(url, tmp, mode = "wb")
    files <- unzip(tmp, list = TRUE)
    readr::read_csv(unz(tmp, files$Name[1]))
  }}()


url <- "https://ed-public-download.scorecard.network/downloads/Most-Recent-Cohorts-Field-of-S

field_of_study <- tempfile(fileext = ".zip") |>
  {\(tmp) {
    download.file(url, tmp, mode = "wb")
    files <- unzip(tmp, list = TRUE)
    readr::read_csv(unz(tmp, files$Name[1]))
  }}()
```

This survey design is a two-stage probability sampling—with states first are selected using probability proportional to size sampling based on the number of eligible colleges. Setting it up this way balances the likelihood of inclusion across large and small states, which helps us when thinking of ideas like number of undergraduates and earnings, which may vary by state size. Then, within each of the ten sampled states, we drew a simple random sample of five colleges is drawn.

```r
math_rows <- field_of_study |>
  filter(startsWith(as.character(CIPCODE), "27"))

college_scorecard_unique <- college_scoreboard |>
  group_by(OPEID6) |>
  summarise(
    INSTNM = first(INSTNM),
    STABBR = first(STABBR),
    CONTROL = first(CONTROL),
    UGDS = case_when(
      all(is.na(UGDS)) ~ NA_real_,
      sum(!is.na(UGDS)) > 1 ~ sum(UGDS, na.rm = TRUE),
      TRUE ~ max(UGDS, na.rm = TRUE)
      ),
    MEDIAN_EARNINGS = mean(MD_EARN_WNE_P10, na.rm = TRUE),
    STAT_MAJOR_IND = max(if_else(!is.na(PCIP27) & PCIP27 > 0, 1, 0), na.rm = TRUE)
)
```

```
Warning: There were 245 warnings in `summarise()`.
The first warning was:
i In argument: `UGDS = case_when(...)`.
i In group 114: `OPEID6 = "001165"`.
Caused by warning in `max()`:
! no non-missing arguments to max; returning -Inf
i Run `dplyr::last_dplyr_warnings()` to see the 244 remaining warnings.
```

```r
state_counts <- college_scorecard_unique |>
  group_by(STABBR) |>
  summarise(num_colleges = n())

n_states <- 10
pi_states <- n_states * state_counts$num_colleges / sum(state_counts$num_colleges)
pi_states[pi_states > 1] <- 1

pps_state_index <- UPsystematic(pi_states)
sampled_states <- state_counts[pps_state_index == 1, ]
sampled_states
```

```
# A tibble: 10 x 2
   STABBR num_colleges
   <chr>         <int>
```

```
 1 AL               70
 2 CA              540
 3 FL              266
 4 IL              208
 5 MD               67
 6 NC              142
 7 NY              380
 8 OK               92
 9 SC               75
10 UT               53
```

```r
colleges_in_sampled_states <- college_scorecard_unique |>
  filter(STABBR %in% sampled_states$STABBR)

set.seed(322)

sampled_colleges <- colleges_in_sampled_states |>
  group_by(STABBR) |>
  slice_sample(n = 5, replace = FALSE)

sampled_colleges <- sampled_colleges |>
  left_join(state_counts, by = "STABBR") |>
  left_join(data.frame(STABBR = state_counts$STABBR, pi_state = pi_states), by = "STABBR") |>
  mutate(
    N_i = num_colleges,
    n_i = 5,
    pi_college_within_state = n_i / N_i,
    weight = 1 / (pi_state * pi_college_within_state)
  )

sampled_colleges_clean <- sampled_colleges |>
  group_by(STABBR) |>
  filter(any(!is.na(UGDS) & !is.nan(UGDS))) |>
  ungroup()

write_csv(sampled_colleges_clean, "college_data.csv")

pps_design <- svydesign(id = ~STABBR + OPEID6, weights = ~weight,
                        data = sampled_colleges_clean)
```

What is an estimate of the total number of undergraduate students enrolled in colleges?

The PPS selection at the first stage ensures that larger states, with more colleges and often also larger colleges (e.g. UCLA in CA, UT Austin in Texas) are more likely to be included to improve the precision of our estimates. The second stage, using simple random sampling within states, provides unbiased representation of institutions within each selected area.

The 95% confidence interval for the estimated total is (4,184,033; 10,273,087) with a point estimate ~7 million full-time enrolled undergraduate students.

```
svytotal(~UGDS, pps_design, na.rm = TRUE)
```

```
        total      SE
UGDS 19716890 5398830
```

```
confint(svytotal(~UGDS, pps_design, na.rm = TRUE))
```

```
       2.5 %   97.5 %
UGDS 9135377 30298402
```

What fraction of colleges have a major in statistical science?

Once again, the PPS selection at the first stage ensures that larger states, which will also offer a variety of academic programs, have a higher probability of being included.

The 95% confidence interval for the estimated proportion is (0.149; 0.524) with a point estimate of about 33.7%.

```
svymean(~STAT_MAJOR_IND, pps_design)
```

```
                  mean    SE
STAT_MAJOR_IND 0.26218 0.074
```

```
confint(svymean(~STAT_MAJOR_IND, pps_design))
```

```
                 2.5 %    97.5 %
STAT_MAJOR_IND 0.11714 0.4072292
```

What is the average of the median earnings among alumni from public schools? From private schools (combine for profit and not for profit private schools)?

Some states with larger college presences also have more jobs within their state. The sample that we took earlier was split into public and private schools.

The estimated average of median earnings among alumni from public colleges is $46,828, with a 95% confidence interval ranging from $37,740 to $55,915. In contrast, alumni from private colleges have a higher estimated average median earning of $60,416, with a wider 95% confidence interval from $43,011 to $77,820. Both of the confidence intervals do cover similar parts for alumni earnings, but roughly our survey suggests graduates from private colleges tend to earn more than those from public colleges. The larger variance in the private colleges seems to be a result of the sampling (and nation) having more public than private colleges.

```
public_design <- subset(pps_design, CONTROL == 1)
svymean(~MEDIAN_EARNINGS, public_design, na.rm = TRUE)
```

```
                 mean     SE
MEDIAN_EARNINGS 46019 3318.9
```

```
confint(svymean(~MEDIAN_EARNINGS, public_design, na.rm = TRUE))
```

```
                   2.5 %   97.5 %
MEDIAN_EARNINGS 39514.38 52524.15
```

```
private_design <- subset(pps_design, CONTROL == 2)
svymean(~MEDIAN_EARNINGS, private_design, na.rm = TRUE)
```

```
                 mean     SE
MEDIAN_EARNINGS 63075 6600.7
```

```
confint(svymean(~MEDIAN_EARNINGS, private_design, na.rm = TRUE))
```

```
                   2.5 %   97.5 %
MEDIAN_EARNINGS 50137.64 76011.77
```

What is the average of the median earnings among alumni from schools with and without stats majors?

For reasons similar to the previous questions answered, we believe the style of sampling we did is well-suited to help us answer this question.

The estimated average of median earnings among alumni from colleges with a statistics major is \$65,266, with a 95% confidence interval ranging from \$52,408 to \$78,125. In contrast, alumni from colleges without a statistics major have an estimated average median earning of \$38,103, with a 95% confidence interval from \$27,572 to \$48,635. Unlike the previous question, these two confidence intervals do not have any overlap. This gives us a pretty good confidence that colleges with statistics majors have higher alumni salaries than colleges without, on average.

```
has_stats <- subset(pps_design, STAT_MAJOR_IND == 1)
svymean(~MEDIAN_EARNINGS, has_stats, na.rm = TRUE)
```

```
                 mean     SE
MEDIAN_EARNINGS 63131 4534.6
```

```
confint(svymean(~MEDIAN_EARNINGS, has_stats, na.rm = TRUE))
```

```
                   2.5 %   97.5 %
MEDIAN_EARNINGS 54242.92 72018.28
```

```
no_stats <- subset(pps_design, STAT_MAJOR_IND == 0)
svymean(~MEDIAN_EARNINGS, no_stats, na.rm = TRUE)
```

```
                 mean     SE
MEDIAN_EARNINGS 38276 3024.6
```

```
confint(svymean(~MEDIAN_EARNINGS, no_stats, na.rm = TRUE))
```

```
                   2.5 %   97.5 %
MEDIAN_EARNINGS 32348.22 44204.28
```

We used ChatGPT on October 18, 2025, to help with writing and debugging sections of the R code for this assignment. The tool assisted us mainly in structuring code for weighted sampling and summary statistics.