STA322 Project 1

Austin Huang and Abby Li

Sampling Data

To sample the data, we implemented a stratified sampling method with proportional allocation.

To start, we found the most up-to-date csv file (Most-Recent-Cohorts-Instition_05192025) from the U.S. Department of Education College Scoreboard site (https://collegescorecard.ed.gov/data/). We decided to create five stratum - private/small, private/medium + large, public/small, public/medium, and public/large. To do this, the first step was to designate which schools were private or public. The variable CONTROL from the original csv file was recoded into a new variable called "type". The variable CONTROL contained 3 levels - 1 indicating public, 2 indicating private nonprofit, and 3 indicating private for profit. The new "type" variable was recoded to contain two levels: public schools (1), and private schools (2 + 3). Next, we categorized the schools as being small, medium, or large. Based on the UGDS variable, indicating undergraduate enrollment, schools were considered small if enrollment is less than 2,000, medium if the enrollment is between 2,000-10,000, and large if enrollment is over 10,000. These thresholds were created based on the size categories from the website. Additionally, we decided to combine the strata for private schools to include both large and medium institutions because there are only ~20 schools that are considered large private schools. This allows us to avoid very small stratum sample sizes when implementing proportional allocation, which can help improve the stability of variance estimates. Overall, we chose to use these 5 strata to create groups based on similar characteristics of type and size with the goal of reducing variance within strata, but also maintaining differences between strata.

We chose to use a stratified random sampling design with proportional allocation, with strata as described above. To implement proportional allocation, we first determined the total number of sampled colleges to be 100. We computed the stratum population sizes N_h. Using the formula for proportional allocation ($n_h = n\frac{N_h}{N}$), we computed the raw n_h for each strata. To round the n_h's to integer while keeping their sum equal to n, we floored each n_h and distributed the remainders to strata with largest fractional parts. After computing the sample size for each strata, we joined these values back into our main dataframe and took SRS without replacement within each strata. We also computed the inclusion probabilities π as well as the weights to use for Horvitz-Thompson unbiased estimators.

```r
library(dplyr)
library(tidyr)
library(survey)

colleges <- read.csv("colleges.csv")
```

```r
colleges_q1 <-  colleges %>%
  select(INSTNM, UGDS, CONTROL, CIP27ASSOC, CIP27BACHL, CIP27CERT1, CIP27CERT2,
         CIP27CERT4, MD_EARN_WNE_5YR, TUITIONFEE_IN) %>%
  mutate (type = case_when(
    CONTROL == 1 ~ "Public",
    CONTROL %in% c(2, 3) ~"Private",
    TRUE ~ NA_character_
  )) %>%
  mutate(size = case_when(
    UGDS > 15000 ~ "Large",
    UGDS >= 2000 & UGDS <= 15000 ~ "Medium",
    UGDS < 2000 ~ "Small",
    TRUE ~ NA_character_
  ))

# creating the six strata
```

```r
colleges_q1 <- colleges_q1 %>%
  mutate(stratum = case_when(
    type == "Public" & size == "Large"   ~ "Public-Large",
    type == "Public" & size == "Medium" ~ "Public-Medium",
    type == "Public" & size == "Small"  ~ "Public-Small",
    type == "Private" & size == "Small" ~ "Private-Small",
    type == "Private" & size %in% c("Medium","Large")~ "Private-Medium/Large",
    TRUE ~ NA_character_
  ))

colleges_q1 <- colleges_q1 %>%
  mutate(has_stats_major = ifelse(
    CIP27CERT1 > 0 | CIP27CERT2 > 0 | CIP27ASSOC > 0 |
    CIP27BACHL > 0 | CIP27CERT4 > 0,
    1, 0
  ))


colleges_q1 <- colleges_q1 %>%
  drop_na(type, size)
```

```r
# sample with proportional allocation
N_total <- nrow(colleges_q1)
n_total <- 100

set.seed(322)

# compute N_h: sample size in each stratum
stratum_info <- colleges_q1 %>%
  group_by(stratum) %>%
  summarise(N_h = n()) %>%
  ungroup()

# compute raw n_h, sample size to allocate to each stratum
stratum_info <- stratum_info %>%
  mutate(n_h_raw = n_total * (N_h / sum(N_h)))

# round n_h to integers while ensuring sum = n_total
stratum_info <- stratum_info %>%
  mutate(n_floor = floor(n_h_raw),
         frac = n_h_raw - n_floor)

remainder <- n_total - sum(stratum_info$n_floor)
if (remainder > 0) {
  # give +1 to the strata with largest fractional parts
  add_idx <- order(stratum_info$frac, decreasing = TRUE)[1:remainder]
  stratum_info$n_final <- stratum_info$n_floor
  stratum_info$n_final[add_idx] <- stratum_info$n_final[add_idx] + 1
} else {
  stratum_info$n_final <- stratum_info$n_floor
}

# If any n_final == 0 (small strata), enforce a minimum of 1:
min_per_stratum <- 1
```

```r
zeros <- which(stratum_info$n_final < min_per_stratum)
if(length(zeros)>0){
  for(i in zeros) stratum_info$n_final[i] <- min_per_stratum
  # reduce from the largest strata to keep sum == n_total
  while(sum(stratum_info$n_final) > n_total){
    # pick stratum with largest n_final to decrement (but keep >= min)
    idx <- which.max(stratum_info$n_final)
    if(stratum_info$n_final[idx] > min_per_stratum)
      stratum_info$n_final[idx] <- stratum_info$n_final[idx] - 1
    else break
  }
}
# check
stopifnot(sum(stratum_info$n_final) == n_total)


# add n_h to colleges_df
colleges_df <- colleges_q1 %>%
  left_join(stratum_info %>%
              select(stratum, N_h, n_final), by = "stratum")
names(colleges_df)[names(colleges_df) == "n_final"] <- "n_h"

# SRSWOR in each stratum
sampled <- colleges_df %>%
  group_by(stratum) %>%
  sample_n(size = unique(n_h)) %>%
  ungroup()

# compute weights - SRSWOR: pi = n_h / N_h
sampled <- sampled %>%
  mutate(pi = n_h / N_h,
         weight = N_h / n_h)
```

Question 1

What is an estimate of the total number of undergraduate students enrolled in colleges? Design your survey with the goal of answering this question most accurately.

```r
des <- svydesign(id = ~1, strata = ~stratum, weights = ~weight, fpc = ~N_h,
                 data = sampled)

total_undergrads <- svytotal(~UGDS, des)
total_undergrads
```

```
##         total      SE
## UGDS 14235630 1223757
```

```r
confint(total_undergrads)
```

```
##         2.5 %   97.5 %
## UGDS 11837110 16634150
```

To answer this question, we use the variable UGDS which indicates the number of undergraduates enrolled in the fall (https://collegescorecard.ed.gov/files/InstitutionDataDocumentation.pdf).

The estimated total number of undergraduate students enrolled in college is 14,235,630 with a standard error of 1,223,757. We are 95% confident that the true total number of undergraduates is between 11,837,110 and 16,634,150.

Question 2

What fraction of colleges have a major in statistical science? Use your sample to estimate this fraction; do not get the answer by filtering the schools.

```
prop_stats <- svymean(~has_stats_major, des)
prop_stats
```

```
##                     mean     SE
## has_stats_major 0.28934 0.0336
```

```
confint(prop_stats)
```

```
##                     2.5 %    97.5 %
## has_stats_major 0.2235079 0.3551808
```

The CIPCODE for mathematics and statistics majors is 27, and there are several degree programs that may be available at each school, including the CIP27ASSOC, CIP27BACHL, CIP27CERT1, CIP27CERT2, CIP27CERT4 variables, which are Boolean variables. If any of these are offered at the college, we count that the university has a statistical science major. The estimated proportion of colleges that have a major in statistical science is 0.2893 with a standard error of 0.0336 The 95% confidence interval for this is (0.2235, 0.3552). (https://collegescorecard.ed.gov/files/FieldOfStudyDataDocumentation.pdf)

Question 3

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)

```
avg_median_earnings <- svyby(~MD_EARN_WNE_5YR, ~type, des, svymean, na.rm = TRUE)
avg_median_earnings
```

```
##            type MD_EARN_WNE_5YR       se
## Private Private        49202.90 2799.480
## Public   Public        49024.78 1459.838
```

```
confint(avg_median_earnings)
```

```
##            2.5 %   97.5 %
## Private 43716.02 54689.78
## Public  46163.55 51886.01
```

To answer this question, we use the MD_EARN_WNE_5YR variable. This variable indicates the median earnings 5 years after graduation, which is a good time metric since the question is focused on alumni specifically (https://collegescorecard.ed.gov/files/InstitutionDataDocumentation.pdf).

The average of the median earnings among alumni 5 years after completion of private colleges is 49202.90 with a standard error of 2799.480. The average of the median earnings among alumni 5 years after completion of public colleges is 49024.78 with a standard error of 1459.838 The 95% confidence interval for median earnings for private school alum is between 43,716.02 and 54,689.78. The 95% confidence interval for median earnings for public school alum is between 46,163.55 and 51,886.01.

Question 4

What is the average of the tuition for in-state students from public schools? From private schools? (combine for profit and not for profit private schools)

```r
avg_tuition <- svyby(~TUITIONFEE_IN, ~type, des, svymean, na.rm = TRUE)
avg_tuition
```

```
##            type TUITIONFEE_IN       se
## Private Private     27193.650 2301.677
## Public   Public      6729.941  916.882
```

```r
confint(avg_tuition)
```

```
##             2.5 %    97.5 %
## Private 22682.446 31704.854
## Public   4932.886  8526.997
```

To answer this question, we use the TUITIONFEE_IN variable. This variable indicates the tuition and required fees that are for in-state students (https://collegescorecard.ed.gov/files/InstitutionDataDocumentation.pdf)

The average of tuition for in-state students in private colleges is 27,193.65 with a standard error of 2,301.68. The average tuition for in-state students in public colleges is 6,729.94 with a standard error of 916.88. The 95% CI for average tuition of in-state students in private colleges is between 22,682.45 and 31,704.85. The 95% CI for average tuition of in-state students in public colleges is between 4,932.89 and 8,527.00. This makes sense, as private colleges usually charge both in-state and out-state students the same tuition, whereas those living in-state and attending public colleges are usually charged significantly less.