Stat 138: Introduction to Sampling Designs
Problem Set 2

Anne Christine Amores

March 21, 2025

Household Energy Use

Suppose that a city has 90,000 dwelling units, of which 35,000 are houses, 45,000 are apartments, and 10,000 are condominiums.

(a) **You believe that the mean electricity usage is about twice as much for houses as for apartments or condominiums, and that the standard deviation is proportional to the mean so that $S_1 = 2S_2 = 2S_3$. How would you allocate a stratified sample of 900 observations if you wanted to estimate the mean electricity consumption for all households in the city?**

To allocate the 900 observations optimally, we use **Neyman allocation**, which accounts for both stratum size and varaibility. Given that the standard deviations are proportional to the means, such that $S_1 = 2S_2 = 2S_3$, the allocation formula we will use is:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} \cdot n$$

where $N_1 = 35,000, N_2 = 45,000, N_3 = 10,000$. Computing for $N_h S_h$ for all three strata,

$$N_1 S_1 = 35,000 \cdot 2S = 70,000S$$

$$N_2 S_2 = 45,000 \cdot S = 45,000S$$

$$N_3 S_3 = 10,000 \cdot S = 10,000S$$

Summing up the $N_h S_h$ for all three strata,

$$\sum_{h=1}^{3} N_h S_h = 70,000S + 45,000S + 10,000S = 125,000S$$

Computing n_h for all three strata:

$$n_1 = \frac{N_1 S_1}{\sum_{h=1}^{3} N_h S_h} \cdot n = \frac{70,000S}{125,000S} \cdot 900 = \boxed{504}$$

$$n_2 = \frac{N_2 S_2}{\sum_{h=1}^{3} N_h S_h} \cdot n = \frac{45,000S}{125,000S} \cdot 900 = \boxed{324}$$

$$n_3 = \frac{N_3 S_3}{\sum_{h=1}^{3} N_h S_h} \cdot n = \frac{10,000S}{125,000S} \cdot 900 = \boxed{72}$$

Thus, the allocation of the stratified sample of 900 observations would be as follows: **504 houses, 324 apartments, and 72 condos.**

(b) **Now suppose that you take a stratified random sample with proportional allocation and want to estimate the overall proportion of households in which energy conservation is practiced. If 45% of house dwellers, 25% of apartment dwellers, and 3% of condominium residents practice energy conservation, what is p for the population? What gain would the stratified sample with proportional allocation offer over an SRS, that is, what is $V_{prop}(\hat{p}_{str})/V_{srs}(\hat{p}_{SRS})$?**

The variance of $\hat{p}_{str}$ is given by

$$V_{(}\hat{p}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_h - n_h}{N_h - 1}\right) \frac{p_h(1-p_h)}{n_h - 1}$$

$$\rightarrow V(\hat{p}_{str}) = \left(\frac{N_1}{N}\right)^2 \left(\frac{N_1 - n_1}{N_1 - 1}\right) \frac{p_1(1-p_1)}{n_1 - 1} + \left(\frac{N_2}{N}\right)^2 \left(\frac{N_2 - n_2}{N_2 - 1}\right) \frac{p_2(1-p_2)}{n_2 - 1} + \left(\frac{N_3}{N}\right)^2 \left(\frac{N_3 - n_3}{N_3 - 1}\right) \frac{p_3(1-p_3)}{n_3 - 1}$$

Setting $n = 900$ then, under proportional allocation, we have the same values for n_1, n_2, and n_3 as in item 1(a).

$$V_{(}\hat{p}_{str}) = \left(\frac{35,000}{90,000}\right)^2 \left(\frac{35,000 - 504}{35,000 - 1}\right) \frac{0.45(0.55)}{504 - 1} + \left(\frac{45,000}{90,000}\right)^2 \left(\frac{45,000 - 324}{45,000 - 1}\right) \frac{0.25(0.75)}{324 - 1} + \left(\frac{10,000}{90,000}\right)^2 \left(\frac{10,000 - 72}{10,000 - 1}\right) \frac{0.03(0.97)}{72 - 1}$$

$$\rightarrow V(\hat{p}_{str}) = 0.0002224513576 \approx 0.000222$$

On the other hand, using SRSWOR, for large populations,

$$V(\hat{p}_{SRS}) = \frac{N-n}{n-1} \frac{p(1-p)}{n}$$

The problem says that 45% of house dwellers, 25% of apartment dwellers, and 3% of condominium residents practice energy conservation. If these proportions can be assumed to be true for the population, then

$$p = \frac{35,000}{90,000}(0.45) + \frac{45,000}{90,000}(0.25) + \frac{10,000}{90,000}(0.03)$$

$$\rightarrow p = \frac{91}{300}$$

$$\rightarrow V(\hat{p}_{SRS}) = \frac{90,000 - 900}{90,000 - 1} \frac{\frac{91}{300}(1 - \frac{91}{300})}{900}$$

$$\rightarrow V(\hat{p}_{SRS}) = 0.0002324570273 \approx 0.000232.$$

Thus,

$$\frac{V_{prop}(\hat{p}_{str})}{V_{srs}(\hat{p}_{SRS})} \approx \frac{0.000222}{0.000232} \boxed{\approx 0.9569}.$$

This tells us that you only need a fraction of the sample size in an SRS to achieve the same precision with a stratified sample. Specifically, only approximately $0.9569n$ observations are needed in a stratified sample using proportional allocation to get the same variance as in a SRSWOR.

Optimizing survey sample size

A public opinion researcher has a budget of \$20,000 for taking a survey. She knows that 90% of all households have telephones. Telephone interviews cost \$10 per household: in-person interviews cost \$30 each if all interviews are conducted in person, and \$40 each if only nonphone households are interviewed in person (because there will be extra travel costs). Assume that the variances in the phone and nonphone groups are similar, and that the fixed costs are $c_0 = \$5000$. How many households should be interviewed in each group if

(a) **all households are interviewed in person**

The \$20,000 budget has to be allocated to the fixed cost and the cost per interview. In this case, since all households will be interviewed in person, the cost per in-person interview is \$30 each. Setting up an equation to determine how many households can be interviewed considering the budget,

$$20,000 = 30n - 500$$
$$30n = 15,000$$
$$n = 500$$

Because of the assumption that the variances in the phone and nonphone groups are similar, we can make use of **proportional allocation**. The problem states that 90% of the households have a phone. It follows that 10% do not. Thus, we multiply 0.9 and 0.1 to the obtained number of households, $n = 500$, to determine how many households are to be interviewed in each group.

Let n_1 be the number of phone households to be interviewed and n_2 be the number of nonphone households to be interviewed. By proportional allocation,

$$n_1 = 500 \cdot 0.9 = \boxed{450}$$
$$n_2 = 500 \cdot 0.1 = \boxed{50}$$

Thus, the researcher can interview 450 phone households and 50 nonphone households, given the budget of \$20,000.

(b) **households with a phone are contacted by telephone and households without a phone are contacted in person**

$$n_h = \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}}} n$$

Since the variances of both groups are assumed to be similar, i.e., $S_1 = S_2 = S$, the equation simplifies to

$$n_h = \frac{\frac{N_h S}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h S}{\sqrt{c_h}}} n$$

$$\to n_h = \frac{\cancel{S} \frac{N_h}{\sqrt{c_h}}}{\cancel{S} \sum_{h=1}^{H} \frac{N_h}{\sqrt{c_h}}} n$$

$$\to n_h = \frac{\frac{N_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h}{\sqrt{c_h}}} n$$

Thus, the equation for the sample size of the stratum corresponding to the phone households is given by

$$n_1 = \frac{\frac{N_1}{\sqrt{c_1}}}{\frac{N_1}{\sqrt{c_1}} + \frac{N_2}{\sqrt{c_2}}} n$$

Since we know that 90% of the households have a phone, then $\frac{N_1}{N} = 0.9 \to N_1 = 0.9N$. In a similar fashion, $\frac{N_2}{N} = 0.1 \to N_2 = 0.1N$. Also, we know that telephone interviews cost \$10 per household and in-person interviews cost \$40 each because only nonphone households will be interviewed in person. Thus, $c_1 = 10$ and $c_2 = 40$. Substituting these into the equation for n_1,

$$n_1 = \frac{\frac{N_1}{\sqrt{c_1}}}{\frac{N_1}{\sqrt{c_1}} + \frac{N_2}{\sqrt{c_2}}} n = \frac{\frac{0.9N}{\sqrt{10}}}{\frac{0.9N}{\sqrt{10}} + \frac{0.1N}{\sqrt{40}}} n$$

$$\to n_1 = \frac{\cancel{N}\frac{0.9}{\sqrt{10}}}{\cancel{N}\left(\frac{0.9}{\sqrt{10}} + \frac{0.1}{\sqrt{40}}\right)} n$$

$$\to n_1 = \frac{18}{19} n.$$

And since there are only two strata,

$$n_2 = n - n_1$$

$$\to n_2 = n - \frac{18}{19} n = n\left(1 - \frac{18}{19}\right) = \frac{1}{19} n.$$

To determine the sample size, n, we use the following equation, which represents the total cost constraint based on the budget and the costs of telephone and in-person interviews:

$$15,000 = 10n_1 + 40n_2$$

$$\to 15,000 = 10\left(\frac{18}{19} n\right) + 40\left(\frac{1}{19} n\right)$$

$$\to 15,000 = n\left(10 \cdot \frac{18}{19} + 40 \cdot \frac{1}{19}\right)$$

$$\to 15,000 = \frac{220}{19} n$$

$$\to n = 1295.454545 \approx 1295,$$

rounded down in consideration of the budget constraint. Then, computing for n_1 and n_2,

$$n_1 = \frac{18}{19} n = \frac{18}{19} \cdot 1295 = 1226.842105 \approx \boxed{1227}$$

$$n_2 = \frac{1}{19} n = \frac{1}{19} \cdot 1295 = 68.15789474 \approx \boxed{68}$$

Thus, if households with a phone are contacted by telephone and households without a phone are contacted in person, the researcher can interview **1228** of the phone households and **68** of the nonphone households.

Trucks

The Vehicle Inventory and Use Survey (VIUS) has been conducted by the U.S. government to provide information on the number of private and commercial trucks in each state. The stratified random sampling design is described in the U.S. Census Bureau (2006b). For the 2002 survey, 255 strate were formed from the sampling frame of truck registrations using stratification variables *state* and *trucktype*. The 50 states plus the District of Columbia formed 51 geographic classes; in each, the truck registrations were partitioned into one of five classes:

1. Pickups
2. Minivans, other light vans, and sport utility vehicles

3. Light single-unit trucks with gross vehicle weight less than 26,000 pounds
4. Heavy single-unit trucks with gross vehicle weight greater than or equal to 26,000 pounds
5. Truck-tractors

Consequently, the full data set has 51 x 5 = 255 strata. Selected variables from the data are in the data file vius.dat. For each question below, give a point estimate and a 95% CI.

(a) **The sampling weights are found in variable *tabtrucks* and the stratification is given by variable *stratum*. Estimate the total number of trucks in the United States. (HINT: What should your response variable be?) Why is the standard deviation of your estimator essentially zero?**

```r
vius_data<-read_excel(path =
                     "C:/Users/amore_6ou078y/OneDrive/Documents/vius.xlsx",
                     col_names=TRUE,
                     trim_ws=TRUE)

head(vius_data)
```

```
## # A tibble: 6 x 27
##   STRATUM ADM_STATE STATE TRUCKTYPE TABTRUCKS HB_STATE BODYTYPE ADM_MODELYEAR
##     <dbl>     <dbl> <chr>     <dbl>     <dbl> <chr>       <dbl>         <dbl>
## 1      11         1 AL            1     3626. AL              1             7
## 2      11         1 AL            1     3626. AL              1            16
## 3      11         1 AL            1     3626. AL              1             5
## 4      11         1 AL            1     3626. AL              1             3
## 5      11         1 AL            1     3626. AL              1            16
## 6      11         1 AL            1     3626. AL              1             5
## # i 19 more variables: VIUS_GVW <dbl>, MILES_ANNL <dbl>, MILES_LIFE <dbl>,
## #   MPG <chr>, OPCLASS <dbl>, OPCLASS_MTR <chr>, OPCLASS_OWN <chr>,
## #   OPCLASS_PSL <chr>, OPCLASS_PVT <chr>, OPCLASS_RNT <chr>, TRANSMSSN <dbl>,
## #   TRIP_PRIMARY <dbl>, TRIP0_50 <chr>, TRIP051_100 <chr>, TRIP101_200 <chr>,
## #   TRIP201_500 <chr>, TRIP500MORE <chr>, ADM_MAKE <dbl>, BUSINESS <dbl>
```

```r
# Computing the total estimate (t_hat)
t_hat <- sum(vius_data$TABTRUCKS, na.rm = TRUE)

# Computing sample size (n_h) and population size (N_h) per stratum
strata_summary <- vius_data %>%
  group_by(STRATUM) %>%
  summarise(n_h = n(),
            N_h = sum(TABTRUCKS, na.rm = TRUE),
            s_h2 = var(TABTRUCKS, na.rm = TRUE))

# Computing the standard deviation
std_error <- sd(vius_data$TABTRUCKS, na.rm = TRUE) / sqrt(nrow(vius_data))

# Computing the stratified variance
var_t_hat <- sum((strata_summary$N_h^2 / strata_summary$n_h) *
    strata_summary$s_h2, na.rm = TRUE)
SE_t_hat <- sqrt(var_t_hat)

# Computing a 95% CI
lower_CI <- t_hat - 1.96 * SE_t_hat
upper_CI <- t_hat + 1.96 * SE_t_hat
```

```r
cat("Estimated total of trucks in the United States:", t_hat, "\n")
cat("Standard error of the estimator:", SE_t_hat, "\n")
cat("95: CI: [", lower_CI, ",", upper_CI, "]\n")
```

```
## Estimated total of trucks in the United States: 85174776
## Standard error of the estimator: 0
## 95: CI: [ 85174776 , 85174776 ]
```

The standard deviation of $\hat{t}$ is essentially zero because the variable *tabtrucks* represents weights for the entire population, not just a sample. And since we are summing these weights, the estimate is already scaled to the full population, leaving little variability in the estimator.

(b) **Estimate the total number of truck miles driven in 2002 (variable *miles_annl*).**

```r
# Estimating the total number of truck miles driven in 2002
t_hat_miles = sum(vius_data$MILES_ANNL * vius_data$TABTRUCKS, na.rm = TRUE)
options(scipen = 999)
cat("Estimated total truck miles driven in 2002:", format(t_hat_miles,
    big.mark = ",", scientific = FALSE), "\n")

# Computing standard error (SE)
se_t_hat <- sqrt(sum((vius_data$MILES_ANNL * vius_data$TABTRUCKS)^2,
    na.rm = TRUE))

# Computing the margin of error for 95% CI
z_alpha <- 1.96
margin_of_error <- z_alpha * se_t_hat

# Computing 95% CI
lower_CI <- t_hat_miles - margin_of_error
upper_CI <- t_hat_miles + margin_of_error


options(scipen = 999)
cat("Estimated total truck miles driven in 2002:", format(t_hat_miles,
    big.mark = ",", scientific = FALSE), "\n")
cat("95% CI: [", format(lower_CI, big.mark = ",", scientific = FALSE),
    ",", format(upper_CI, big.mark = ",", scientific = FALSE), "]\n")
```

```
## Estimated total truck miles driven in 2002: 1,114,727,883,443
## 95% CI: [ 1,092,865,546,984 , 1,136,590,219,902 ]
```

(c) **Estimate the total number of truck miles driven in each of the five *trucktype* classes.**

```r
# Estimating the total number of truck miles driven in each of the five
trucktype classes
total_miles_by_trucktype <- vius_data %>%
  group_by(TRUCKTYPE) %>%
  summarise(estimated_total_miles = sum(MILES_ANNL * TABTRUCKS,
    na.rm = TRUE)) %>%
  ungroup()


# Adding a column for the 95% CIs
total_miles_by_trucktype <- total_miles_by_trucktype %>%
```

```
  mutate(
    lower_CI = estimated_total_miles - (1.96 * (sd(estimated_total_miles) /
        sqrt(n()))),
    upper_CI = estimated_total_miles + (1.96 * (sd(estimated_total_miles) /
        sqrt(n())))
  )

print(total_miles_by_trucktype)
```

```
## # A tibble: 5 x 4
##   TRUCKTYPE estimated_total_miles lower_CI      upper_CI
##       <dbl>                 <dbl>    <dbl>         <dbl>
## 1         1         428294502082.  2.16e11 641043144761.
## 2         2         541099850893.  3.28e11 753848493572.
## 3         3          41279084490. -1.71e11 254027727169.
## 4         4          31752656137. -1.81e11 244501298816.
## 5         5          72301789843. -1.40e11 285050432522.
```

(d) **Estimate the average miles per gallon (MPG) for the trucks in the population.**

```
# Estimating the average miles per gallon (MPG) for the trucks in the
population
vius_data$MPG <- as.numeric(vius_data$MPG)
avg_mpg <- (sum(vius_data$MPG * vius_data$TABTRUCKS,
    na.rm = TRUE) / sum(vius_data$TABTRUCKS, na.rm = TRUE))
# Computing standard deviation and sample size
sd_mpg <- sd(vius_data$MPG, na.rm = TRUE)
n_mpg <- sum(!is.na(vius_data$MPG))

# Computing standard error (SE) and 95% CI
se_mpg <- sd_mpg / sqrt(n_mpg)
lower_ci <- avg_mpg - (1.96 * se_mpg)
upper_ci <- avg_mpg + (1.96 * se_mpg)
cat("Average miles per gallon (MPG) for the trucks in the population:",
    avg_mpg, "\n")
cat("95% CI:[", lower_ci, ",", upper_CI, "]\n")
```

```
## Average miles per gallon (MPG) for the trucks in the population: 13.43274
## 95% CI:[ 13.39429 , 1136590219902 ]
```

Baseball data

Exercise 32 of Chapter 2 described the population of baseball players in data file baseball.dat.

(a) **Take a stratified random sample of 150 players from the file, using proportional allocation with the different teams as strata. Describe how you selected the sample.**

Importing the dataset,

```
baseballdata<-read_excel(path =
            "C:/Users/amore_6ou078y/OneDrive/Documents/baseball.xlsx",
            col_names=c("team", "leagueID", "player", "salary", "POS", "G",
            "GS", "InnOuts", "PO", "A", "E", "DP", "PB", "GB", "AB", "R",
            "H", "SecB", "ThiB", "HR", "RBI", "SB", "CS", "BB", "SO",
            "IBB", "HBP", "SH", "SF", "GIDP"),
```

```
            trim_ws=TRUE)
head(baseballdata)
```

```
## # A tibble: 6 x 30
##   team  leagueID player salary POS       G    GS InnOuts    PO     A
##   <chr> <chr>    <chr>    <dbl> <chr> <dbl> <dbl>   <dbl> <dbl> <dbl>
## 1 ANA   AL       ander~ 6.20e6 CF      112    92    2375   211     5
## 2 ANA   AL       colon~ 1.10e7 P         3    34     625     8    30
## 3 ANA   AL       davan~ 3.75e5 CF      108    27     743    75     1
## 4 ANA   AL       donne~ 3.75e5 P         5     0     126     2     2
## 5 ANA   AL       eckst~ 2.15e6 SS      142   136    3575   198   309
## 6 ANA   AL       ersta~ 7.75e6 1B      125   124    3196   986    66
## # i 20 more variables: E <dbl>, DP <dbl>, PB <chr>, GB <dbl>,
## #   AB <dbl>, R <dbl>, H <dbl>, SecB <dbl>, ThiB <dbl>, HR <dbl>,
## #   RBI <dbl>, SB <dbl>, CS <dbl>, BB <dbl>, SO <dbl>, IBB <chr>,
## #   HBP <chr>, SH <dbl>, SF <chr>, GIDP <dbl>
```

```
# Computing the number of players per team (Nh)
Nh <- baseballdata %>% count(team, name = "Nh")

# Computing the standard deviation of salary per team (Sh)
Sh <- baseballdata %>% group_by(team) %>% summarise(Sh = sd(salary,
    na.rm = TRUE)) %>% ungroup()

# Merging Nh and Sh into one data frame and dropping rows w/ missing values
NhSh <- left_join(Nh, Sh, by = "team") %>% drop_na(Sh)

# Defining the sample sizes based on proportional allocation
total_sample_size <- 150
NhSh <- NhSh %>%
    mutate(sample_size = round(Nh / sum(Nh) * total_sample_size))

# Sorting data by the stratification variable (team)
baseballdata <- arrange(baseballdata, team)

# Drawing a stratified sample using SRSWOR
set.seed(138)
strat_sample <- strata(baseballdata, stratanames = c("team"),
    size = NhSh$sample_size, method = "srswor")

# Extracting the sampled data
baseball_sample <- getdata(baseballdata, strat_sample)

head(baseball_sample)
```

```
##    leagueID   player   salary POS   G  GS InnOuts  PO  A E DP PB  GB  AB   R
## 10       AL greggke0   301500   P   5   0     263   2  5 0  1  .   5   0   0
## 11       AL guerrvl0 11000000  RF 156 143    3702 308 13 9  2  . 156 612 124
## 16       AL molinbe0  2025000   C  97  89    2286 597 56 3  5  6  97 337  36
## 22       AL salmoti0  9900000  RF  60   5     117  15  1 0  0  .  60 186  15
## 25       AL washbja0  5450000   P   3  25     448   3 22 1  2  .   3   5   0
## 27       NL alomaro0  1000000  2B  38  23     610  48 53 3 10  .  38 110  14
##      H SecB ThiB HR RBI SB CS BB SO IBB HBP SH SF GIDP team ID_unit      Prob
## 10   0    0    0  0   0  0  0  0  0   0   0  0  0    0  ANA      10 0.1923077
```

```
## 11 206    39     2 39 126 15   3 52 74   14    8  0  8    19  ANA        11 0.1923077
## 16   93    13     0 10  54  0   1 18 35    1    2  2  4    18  ANA        16 0.1923077
## 22   47     7     0  2  23  1   0 14 41    0    2  0  4     2  ANA        22 0.1923077
## 25    2     0     0  0   1  0   0  0  0    0    0  2  0     0  ANA        25 0.1923077
## 27   34     5     2  3  16  0   2 12 18    0    1  2  0     2  ARI        27 0.1785714
##     Stratum
## 10        1
## 11        1
## 16        1
## 22        1
## 25        1
## 27        2
```

As seen in the block of code above, a stratified random sample of 150 players was selected using proportional allocation, with teams as strata. First, the number of players per team (N_h) and the standard deviation of salaries (S_h) were computed. Then, the sample size for each team was determined based on its proportion in the total population. The dataset was sorted by team, and a simple random sample without replacement (SRSWOR) was drawn within each stratum to ensure proper representation.

(b) **Find the mean of the variable *logsal*, using your stratified sample, and give a 95% CI.**

```r
# Adding logsal column
baseball_sample <- baseball_sample %>% mutate (logsal = log(salary))

# Computing mean and standard error
logsal_mean <- mean(baseball_sample$logsal, na.rm = TRUE)
logsal_sd <- sd(baseball_sample$logsal, na.rm = TRUE)
n <- nrow(baseball_sample)

# Computing 95% CI
t_value <- qt(0.975, df = n - 1)
margin_error <- t_value * (logsal_sd / sqrt(n))
lower_CI <- logsal_mean - margin_error
upper_CI <- logsal_mean + margin_error

cat("Mean of logsal:", logsal_mean, "\n")
cat("95% CI: [", lower_CI, ",", upper_CI, "]\n")
```

```
## Mean of logsal: 13.82987}
## 95% CI: [ 13.62726 , 14.03249 ]
```

(c) **Estimate the proportion of players in the data set who are pitchers, and give a 95% CI.**

```r
# Estimating the proportion of pitchers
p_hat <- mean(baseball_sample$POS == "P", na.rm = TRUE)

# Sample size

n <- nrow(baseball_sample)

# Population size

N <- 797

# Getting the standard error with FPC
```

```r
SE <- sqrt((1 - (n/N)) * (p_hat * (1- p_hat)) / (n-1))

# Computing 95% CI
z_alpha <- 1.96
lower_CI <- p_hat - z_alpha * SE
upper_CI <- p_hat + z_alpha * SE

cat("Estimated proportion of pitchers in the population:", p_hat, "\n")
cat("95% CI: [", lower_CI, ",", upper_CI, "]\n")
```

```
## Estimated proportion of pitchers in the population: 0.5
## 95% CI: [ 0.4276638 , 0.5723362 ]
```

(e) **Examine the sample variances in each stratum. Do you think optimal allocation would be worthwhile for this problem?**

```r
# Computing sample variances in each stratum
sample_variances <- baseball_sample %>%
  group_by(team) %>%
  summarise(variance_logsal = var(logsal, na.rm = TRUE))
head(sample_variances)
```

```
## # A tibble: 6 x 2
##   team  variance_logsal
##   <chr>           <dbl>
## 1 ANA             2.22
## 2 ARI             0.310
## 3 ATL             3.45
## 4 BAL             0.175
## 5 BOS             2.69
## 6 CHA             0.205
```

It seems that proportional allocation wouldn't work well for this problem because it assumes similar variances across strata, and in this case, the sample variances of *logsal* differ significantly across teams, ranging from roughly 0.004 to over 3.4. **Optimal allocation is worthwhile for this problem,** as it is better to use when variances differ significantly and it gives more samples to high-variance teams, improving estimate precision.