Homework 1

Kevin Linares

2025-01-23

Download the Excel file "fem1524_admin.xlsx" from the homework folder on the course Canvas site. This file is a population list of N = 2,920 young women between the ages of 15 and 24, which will be considered as a sampling frame for this first homework assignment.

1. Select a simple random sample (SRS) of size n = 20 from this frame. Each student will select a different simple random sample, using the R code set.seed(the last four digits of your UM/UMD student ID). Note that we are simulating the notion of hypothetical repeated random sampling using the same SRS design! The class has 30 enrolled students and would generate 30 samples.

```
fem_dat <- read_xlsx(
  "~/repos/UMD_classes_code/applied_sampling_SURV625/homework/fem1524_admin.xlsx"
  )

glimpse(fem_dat)
```

```
Rows: 2,920
Columns: 4
$ AGER    <dbl> 23, 24, 24, 22, 19, 24, 23, 17, 15, 18, 18, 15, 24, 24, 21, 22~
$ cluster <dbl> 1052, 1052, 1052, 1052, 1052, 1172, 1052, 1172, 1172, 1172, 11~
$ stratum <dbl> 105, 105, 105, 105, 105, 117, 105, 117, 117, 117, 117, 105, 10~
$ ID      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,~
```

```
# random sample of size N.
set.seed(4291)
fem_dat_sample <- fem_dat |> sample_n(size=20)
```

2. Give, in selection order, the list of the 20 four-digit selection number (IDs) and the values of AGE for the women in your sample.

```
# confirm that the same 20 IDs were selected due to seed
fem_dat_sample |>
  mutate(selection = row_number()) |>
  select(selection, ID, AGER) |>
  print(n=20)
```

```
# A tibble: 20 x 3
   selection    ID  AGER
       <int> <dbl> <dbl>
 1         1  1954    21
 2         2  2009    21
 3         3  1698    18
 4         4   370    18
 5         5    82    21
 6         6  2135    21
 7         7   318    21
 8         8  1702    24
 9         9  2188    20
10        10   265    22
11        11   822    23
12        12   157    19
13        13  2660    21
14        14    33    18
15        15  2856    22
16        16   863    21
17        17  2330    21
18        18  1060    22
19        19   575    19
20        20   884    20
```

3. Compute the sample estimate of the mean age. What else would we need to compute (be specific) to make inference about the mean age of the population?

    - The sample mean is given as:

$$\bar{y} = \frac{1}{N} \sum_{i \epsilon s} y_i$$

```
ave_age <- fem_dat_sample |>
  summarize(ave_age =
              # divide sum of each age value by sample size
              sum(AGER)/n()
            ) |> pull()

print(ave_age)
```

```
[1] 20.65
```

- The sample mean is 20.6. To make an inference about the population age mean we would need to compute the finite population correction factor, in our case we can assume that the sample size is the population size of N=2920 and which we plug into our equation $f = n/N = 20/2920 = .007$.

```
f <- nrow(fem_dat_sample) / nrow(fem_dat); f
```

```
[1] 0.006849315
```

- After calculating the finite population correction factor we can use it to compute the element variance estimate $s^2 = \frac{1}{n-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$ as;

```
# estimate s^2 using fcp
elem_var_est <- fem_dat_sample |>
  # apply s^2 equation
  summarise( s_squared =
               (1 / (n() - 1) ) * (
    sum( (AGER - mean(AGER))^2) )
  ) |> pull()

# estimate var(y_hat)
samp_var_est <- fem_dat_sample |>
  summarise( var_y_bar =
               (1 - f) * (elem_var_est / n())
             ) |> pull()

# estimate standard error
se_y_bar <- sqrt(samp_var_est)
```

```r
# use the standard error to compute 95% CI
n <- nrow(fem_dat_sample) # sample size
qt_value <- .975 # quantile function to use

Mean_CI <- c(ave_age - qt(qt_value, n-1) * se_y_bar,
             ave_age + qt(qt_value, n-1) * se_y_bar)

# add average age with CI

Mean_CI <- append(ave_age, Mean_CI) |> unlist() |> round(3)


# label CI
names(Mean_CI) <- c("mean", "lower", "upper")

print(Mean_CI |> round(2))
```

```
 mean lower upper
20.65 19.89 21.41
```

- After calculating the element variance estimate $s^2 = 2.6605263$, we can now compute the next element we need which is the sampling variance estimate $var(\bar{y}) = (1-f)\frac{s^2}{n} = 0.1321152$ which we can than use to calculate the standard error $se(\bar{y}) = \sqrt{var(\bar{y})} = 0.3634765$ to be used to compute compute 95% confidence intervals, which indicate the accuracy of an estimate. If we were to take samples from the same population and construct a confidence interval, we would expect 95% of the resulting intervals to include the true value of the population parameter. Our 95% CI is $\bar{y} \pm t_{1-\alpha/2,n-1} \times se(\bar{y})$ in a simple random sample of 20 females we estimate the population average age to be 20.7 (95% CI [19.9., 21.4]).

4. What would we call the distribution that we would see if we plotted all 30 sample estimates of the mean age (computed from the 30 unique samples generated by the students in the class)? What would we call the standard deviation of this distribution?

- We call this the sampling distribution of the age mean, which is the distribution of different values of the statistic obtained by the process of taking all possible samples from the population. **Specifically, the standard deviation** of the sampling distribution is a measure of variability, that tells us how much sample means vary, which we call this standard error, which is an estimate of precision of the sample mean is the standard deviation of the sampling distribution of the mean.

5. Based on the ID numbers of the SRS sample that you selected above, use the data file available for this homework "SM 625 HW 1.xlsx" and work on the following questions.

```
# read new data in
sm_dat <- read_xlsx(
  "~/repos/UMD_classes_code/applied_sampling_SURV625/homework/SM_625_HW_1.xlsx")

# combine with sample
fem_dat_sample_combined <- fem_dat_sample |>
  left_join(sm_dat)
```

a. Look up the number of male sexual partners in the past year (PARTS1YR) that were reported in a survey by each of your 20 selections in the Excel file. Estimate the mean number of partners in the past year for the population, $\bar{y} = \frac{y}{N} = \sum_{i=1}^{20} y_i/n$.

```
# print variable counts
fem_dat_sample_combined |> select(ID, PARTS1YR)
```

```
# A tibble: 20 x 2
      ID PARTS1YR
   <dbl>    <dbl>
 1  1954        1
 2  2009        1
 3  1698        2
 4   370        1
 5    82        0
 6  2135        1
 7   318        1
 8  1702        1
 9  2188        5
10   265        1
11   822        0
12   157        1
13  2660        1
14    33        2
15  2856        1
16   863        1
17  2330        1
18  1060        1
```

```
19   575         1
20   884         0
```

```
y_bar <- fem_dat_sample_combined |>
  summarise(
    ave_sexual_partner = sum(PARTS1YR)/n() # compute mean
  )

y_bar
```

```
# A tibble: 1 x 1
  ave_sexual_partner
               <dbl>
1               1.15
```

- We compute the average sexual partners in the past year from our sample of 20 to be estimated at 1.15.

b. Estimate the population element variance s^2

$$s^2 = \frac{\sum_{i=1}^{20}(y_i - \bar{y})^2}{n-1} = \frac{(\sum_{i=1}^{20} y_i^2 - \frac{y^2}{n})}{n-1}$$

```
s_sqrd <- fem_dat_sample_combined |>
  summarise(
    sum( (PARTS1YR - mean(PARTS1YR))^2) / (n()-1)
    ) |> pull()
s_sqrd
```

```
[1] 1.081579
```

- The population element variance s^2 for sexual partners in the past year is 1.08.

c. Estimate the sampling variance of the mean, $var(\bar{y})$, and the standard error $SE\bar{y}$ as:

$$var(\bar{y}) = (1-f)\frac{s^2}{n}, SE(\bar{y}) = \sqrt{var(\bar{y})}$$

```
# compute f wih sample size
f <- nrow(fem_dat_sample_combined) / nrow(fem_dat)

# compute variance
var_bar_y <- (1 - f) * (s_sqrd / nrow(fem_dat_sample_combined))

# compute standard error
se_bar_y <- sqrt(var_bar_y)
```

- The sampling variance of the mean is 0.05 with a standard error of 0.23.

d. Compute a 95% confidence interval for the sample mean.

$$\bar{y} \pm t_{1-\alpha/2,n-1} \times se(\bar{y})$$

```
# use the standard error to compute 95% CI
n <- nrow(fem_dat_sample_combined) # sample size
qt_value <- .975 # quantile function to use

# compute CI
Mean_CI <- c(y_bar - qt(qt_value, n-1)*se_bar_y,
             y_bar + qt(qt_value, n-1)*se_bar_y)


# add average age with CI
Mean_CI <- append(y_bar, Mean_CI) |> unlist() |> round(3)

# label CI
names(Mean_CI) <- c("mean", "lower", "upper")

# add average age & print
print(Mean_CI |> round(4))
```

```
 mean lower upper
1.150 0.665 1.635
```

- The average number of sexual partners in the past year is 1.15 (95% CI [.67, 1.64]).

e. Explain why the mean computed in a) will generally not be equal to the population mean.

- When we draw a sample from the population, we are typically taking a subset of the population with some inherent randomness. Each sample are slightly different thus leading to variation in the estimate, and this is called sampling variation. Sampling error also contributes to this discrepancy, which is the difference between the sample mean and population mean, and it is unavoidable due to the random nature of sampling from a population. As the sample size increase, the distribution of the sample means approximates the population mean due to the central limit theorem, which states that more samples eventually converges to a normal distribution.

f. Estimate the coefficient of variation of the mean, $CV(\bar{y}) = se(\bar{y})/\bar{y}$.

```
CV <- se_bar_y/y_bar |> pull()
```

- The coefficient variation of the mean for this estimate is 0.202.

g. What difference would it make for the sampling variance of the mean if the sample size were increased to n = 60?

- By increasing the sample to an n of 60, the sampling variances will likely get smaller, hence the standard deviation naturally gets smaller as well. This is because larger samples begin to approximate the population's representation. Furthermore, with more observations the impact of outliers are minimized, leading to more stable and reliable sample means, hence lower variability.

h. What sample size is needed to obtain $se(\bar{y}) = 0.05$? What about a $cv(\bar{y}) = 0.10$? What about a 95% confidence interval with width 0.40 (using 2 for the t-value)?

- If the desired $se(\bar{y}) = 0.05$ we can use the equation to solve for n_0 using $s = \sqrt{s_2}$, a critical value of 1.96, and we assumpe the population size is 2920 to estimate the desired sample size n as:

$$n_0 = (z^2_{\alpha/2} S^2), n = \frac{n_0}{1 + \frac{n_0}{N}}$$

```
se <- .05

# use s_sqrt calculated earlier, take the standard deviation
s <- sqrt(s_sqrd)

# critical value
z <- 1.96

# build equation to calculate n_0
n_0 <- ( ( (z/2) / s) / se )^2

n <- n_0 / ( 1 + (n_0/nrow(fem_dat)) )

n_0
```

```
[1] 355.1844
```

```
n
```

```
[1] 316.6657
```

- When we desire a $se(\bar{y}) = .05$ we would need $n = 317$.

- If we desire a $cv(\bar{y}) = 0.10$, we can use the sample mean estimated earlier to solve for the standard error $cv(\bar{y}) = se/\bar{y} = .10$ and use the equation above.

```
se <- .115


# use s_sqrt calculated earlier, take the standard deviation
s <- sqrt(s_sqrd)

# critical value
z <- 1.96

# build equation to calculate n_0
```

```
n_0 <- ( ( (z/2) / s) / se )^2

n <- n_0 / ( 1 + (n_0/nrow(fem_dat)) )

n_0
```

```
[1] 67.14261
```

```
n
```

```
[1] 65.63344
```

- We note that the sample size needed to achieve a $cv(\bar{y}) = 0.10$ is $n = 67$

- When we desire a 95% CI with a width of .40 we need to use a standard error of .10, and achieve a desired $n = 86$.

```
# theoretical n size
n <- 1000

se <- .10

# use the standard error to compute 95% CI
t <- 2 # quantile function to use

# compute CI
Mean_CI <- c(y_bar - t*se,
             y_bar + t*se)


# add average age with CI
Mean_CI <- append(y_bar, Mean_CI) |> unlist() |> round(3)

# label CI
names(Mean_CI) <- c("mean", "lower", "upper")

# add average age & print
range <- as.numeric(Mean_CI[3] - Mean_CI[2] )
range
```

```
[1] 0.4
```

```
s <- sqrt(s_sqrd)

# critical value
z <- 1.96

# build equation to calculate n_0
n_0 <- ( ( (z/2) / s) / se )^2

n <- n_0 / ( 1 + (n_0/nrow(fem_dat)) )

n_0
```

```
[1] 88.79611
```

```
n
```

```
[1] 86.17554
```

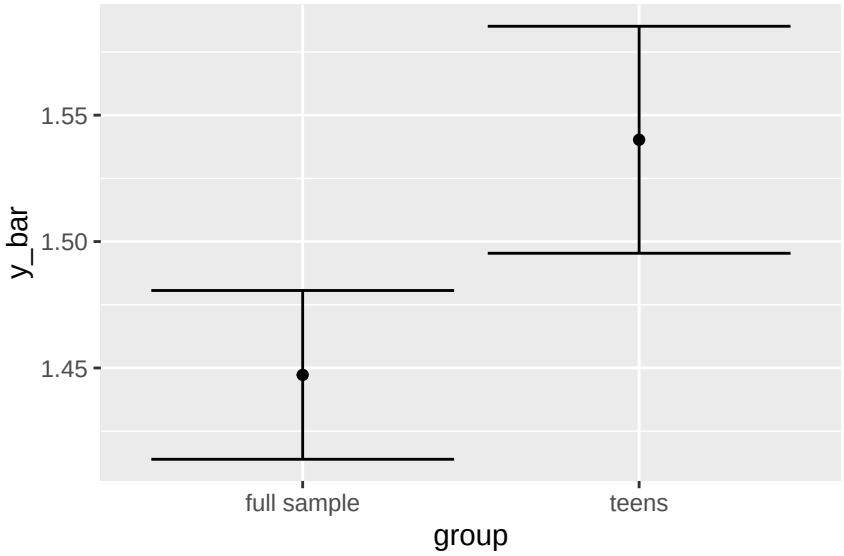
i. Estimate the mean number of male sexual partners in the past year (and its standard error) for the subclass of teenagers (age 15-19) in the sample. Ignore the finite population correction in the calculation of the standard error. How does this standard error compare to the standard error for the full sample? Would you expect such a difference? If so, why?

```
# 1st, combine datasets
dat_combined <- fem_dat |>
  left_join(sm_dat)

# 2nd, calculate y_bar and SE for subset and full sample
full_sample_est <- dat_combined |>
   summarise(y_bar = mean(PARTS1YR),
             se_bar_y = sd(PARTS1YR)/sqrt(n())
             ) |>
  mutate(group="full sample") |>
  select(group, y_bar, se_bar_y)
```

```r
teen_est <- dat_combined |>
  filter(between(AGER, 15, 19)) |>
  summarise(y_bar = mean(PARTS1YR),
            se_bar_y = sd(PARTS1YR)/sqrt(n())

            ) |>
  mutate(group = "teens") |>
  select(group, y_bar, se_bar_y)

# combine estimates
dat <- full_sample_est |>
  add_row(teen_est)

dat
```

```
# A tibble: 2 x 3
  group       y_bar se_bar_y
  <chr>       <dbl>    <dbl>
1 full sample  1.45   0.0334
2 teens        1.54   0.0449
```

```r
# plot estimates
dat |>
  ggplot(aes(x= group, y=y_bar)) +
  geom_point() +
  geom_errorbar(aes(ymin=y_bar - se_bar_y, ymax = y_bar + se_bar_y))
```



- I would expect the standard error of the full sample to be smaller than the sub-sample due to more observations, almost 900 more. However, the teenagers represent a large portion of the sample which perhaps this is why the standard errors are so close between the subset of teen and the full sample.