Homework 1

Kevin Linares

2025-01-21

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(5274)
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.

```r
# 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  1830    19
 2         2  1773    18
 3         3  2024    24
 4         4  2184    23
 5         5  1693    21
 6         6  2606    18
 7         7  1270    19
 8         8   809    23
 9         9  2492    18
10        10  2614    19
11        11  2394    21
12        12    75    18
13        13   483    16
14        14  1891    20
15        15  2487    17
16        16  2276    24
17        17  2185    23
18        18  1525    23
19        19  2538    23
20        20  2093    24
```

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()
            )

print(ave_age)
```

```
# A tibble: 1 x 1
  ave_age
    <dbl>
1    20.6
```

- On it's own the sample mean of 20.62 is not useful for making an inference about the population mean, yet without knowing some information about the spread of values in the sample. We can compute the element variance estimate s^2 and than we can square it to compute the standard deviation $s = \sqrt{s^2}$ as:

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

```
var_est <- fem_dat_sample |>
  summarize(
    var_est =
      (1 / (n()-1)) * sum( (AGER - mean(AGER))^2)
  )

stand_dev <- sqrt(var_est)

print(var_est)
```

```
# A tibble: 1 x 1
  var_est
    <dbl>
1    7.00
```

```
print(stand_dev)
```

```
# A tibble: 1 x 1
  var_est
    <dbl>
1    2.65
```

- We can now use our $s^2 = 2.66$ estimate (with $s = 1.63$) to compute the sampling variance estimate $var(\bar{y})$, which we use a sample to estimate the variances in the population. However, we need to compute the finite population $f = n/N$ which if we assume the overall sample is the population our estimate is $f = 20/2920$, and express the sampling variance estimate as:

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

```
f <- 20/2920
sample_var = (1 - f) * (var_est / nrow(fem_dat_sample))

print(sample_var)
```

```
    var_est
1 0.3474721
```

- We can now use the $var(\bar{y}) = .13$ estimate to compute a standard error $se(\bar{y}) = \sqrt{var(\bar{y})}$ that we will be able to use to 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. We express confidence intervals as:

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

```
se <- sqrt(sample_var) # standard error
n <- nrow(fem_dat_sample) # sample size
qt_value <- .975 # quantile function to use

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

# label CI
```

```r
names(Mean_CI) <- c("lower", "upper")

# add average age & print
append(ave_age, Mean_CI) |> unlist() |> round(3)
```

```
ave_age   lower   upper
 20.550  19.316  21.784
```

- In a simple random sample of 20 females we infer on the average female age in the population We estimate the population average age to be 20.6 (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 error of the 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  1830        1
 2  1773        3
 3  2024        2
 4  2184        0
 5  1693        1
 6  2606        1
 7  1270        1
 8   809        4
 9  2492        1
10  2614        3
11  2394        1
12    75        1
13   483        1
14  1891        1
15  2487        1
16  2276        2
17  2185        0
18  1525        1
```

```
19   2538         6
20   2093         1
```

```r
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                1.6
```

- 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

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

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

s_sqrd
```

```
# A tibble: 1 x 1
  s_sqrd
   <dbl>
1   2.04
```

- The population element variance 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:

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

```r
# compute f wih sample size
f <- 20/2920

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

# compute standard error
se_bar_y <- sqrt(var_bar_y)

var_bar_y |> rename(var_bar_y=1) # variance
```

```
  var_bar_y
1 0.1014059
```

```r
se_bar_y |> rename(se_bar_y=1) # standard error
```

```
   se_bar_y
1 0.3184429
```

- The sampling variance of the mean is .05 with a standard error of .23.

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

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

```r
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)

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

# add average age & print
append(y_bar, Mean_CI) |> unlist() |> round(3)
```

```
ave_sexual_partner              lower              upper
             1.600              0.933              2.267
```

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

CV |> rename(CV=1)
```

```
         CV
1 0.1990268
```

- The coefficient variation of the mean for this estimate is .20.

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)?

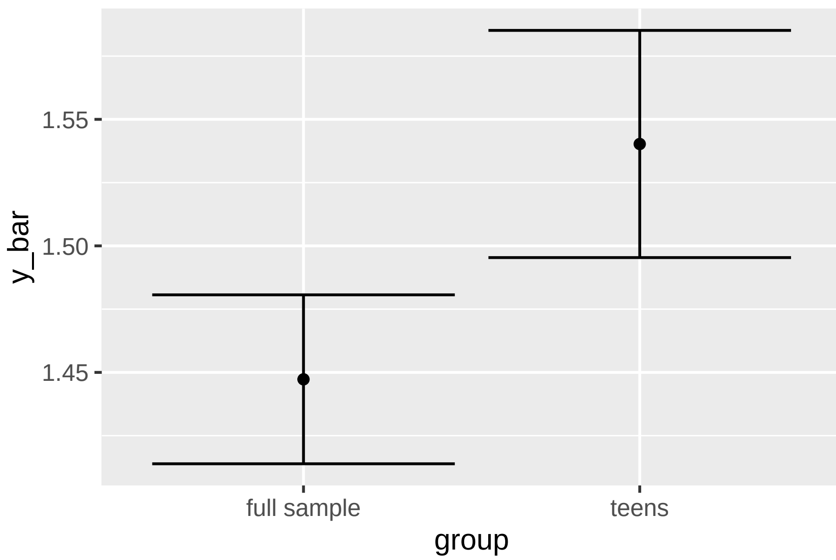
- If the $se(\bar{y}) = 0.05$ than the desired sampling variance is $Var(\bar{y}) = SE(\bar{y}) = .05^2 = .0025$, so we would need a sample of 638 to go from a current SE of .23 to a desired SE of .05.
- We need a sample size of 60 to achieve a CV from the current .20 to .10.
- We need a sample of 94 to go from a .97 95% CI width of .97, to a desired width of .40.

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?

```
dat_combined <- fem_dat |>
  left_join(sm_dat)

full_sample_est <- dat_combined |>
   summarise(y_bar = mean(PARTS1YR),
            s_sqrd =sum( (PARTS1YR-mean(PARTS1YR))^2/(n()-1)),
            var_bar_y =  s_sqrd / n(),
            se_bar_y = sqrt(var_bar_y)
            ) |>
  mutate(group="full sample") |>
  select(group, y_bar, se_bar_y)


teen_est <- dat_combined |>
  filter(between(AGER, 15, 19)) |>
  summarise(y_bar = mean(PARTS1YR),
            s_sqrd =sum( (PARTS1YR-mean(PARTS1YR))^2/(n()-1)),
            var_bar_y =  s_sqrd / n(),
            se_bar_y = sqrt(var_bar_y)
            ) |>
  mutate(group = "teens") |>
  select(group, y_bar, se_bar_y)

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
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. However, the standard errors are not that different, meaning that the precision of these estimates for these two groups are almost comparable. The means are different and what I would expect since older women are probably only have one sexual partner if married.