Homework 1

Kevin Linares

2025-01-17

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_adm

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.

```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  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:

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

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

```r
             )

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:

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

```r
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    2.66
```

```r
print(stand_dev)
```

```
# A tibble: 1 x 1
  var_est
    <dbl>
1    1.63
```

- 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 0.1321152
```

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

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

```
ave_age   lower   upper
 20.650  19.889  21.411
```

- 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]).