SURV625, HW-1

Sagnik Chakravarty

Table of contents

List of Tables

```
library(survey)
library(tidyverse)
library(sampling)
library(SDAResources)
library(dplyr)
library(ggplot2)
library(readxl)
library(knitr)
library(kableExtra)
```

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.

Data:

```
fem1524 <- read_excel('fem1524_admin.xlsx')

knitr::kable(head(fem1524, 5),
             format = 'latex',
             booktabs = TRUE,
             caption = 'Fem1524')%>%
  kable_styling(latex_options = c("hold_position"))
```

Table 1: Fem1524

AGER	cluster	stratum	ID
23	1052	105	1
24	1052	105	2
24	1052	105	3
22	1052	105	4
19	1052	105	5

```
print(dim(fem1524), format = 'pdf')
```

```
[1] 2920    4
```

Question 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.

Code:

```
set.seed(6264) # My UMD id is 121176264

# indexing 20 sample from 2920 population size
index <- srswor(n = 20, N = nrow(fem1524))

print(index[1:10], format = 'pdf')
```

```
 [1] 0 0 0 0 0 0 0 0 1 0
```

```
print(which(index == 1), format = 'pdf')
```

```
 [1]    9  579  821  868  882  907 1145 1215 1254 1355 1718 1891 1944 2037 2283
[16] 2345 2372 2493 2638 2750
```

```
# Selecting the sample from the population
fem1524_sample <- getdata(fem1524, index)

kable(fem1524_sample,
      format = 'latex',
      caption = 'The 20 selected sample')%>%
  kable_styling(latex_options = c("hold_position"))
```

Table 2: The 20 selected sample

ID_unit	AGER	cluster	stratum	ID
9	15	1172	117	9
579	22	1101	110	579
821	23	1311	131	821
868	23	1311	131	868
882	17	1312	131	882
907	19	1271	127	907
1145	18	1382	138	1145
1215	21	1393	139	1215
1254	22	1401	140	1254
1355	19	1404	140	1355
1718	22	1434	143	1718
1891	20	1453	145	1891
1944	24	1461	146	1944
2037	16	1463	146	2037
2283	18	1481	148	2283
2345	21	1492	149	2345
2372	20	1493	149	2372
2493	24	1511	151	2493
2638	16	1531	153	2638
2750	18	1542	154	2750

Question 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.

Code:

```
fem1524_sample_id_age <- fem1524_sample[c('ID', 'AGER')]
fem1524_sample_id_age$SelectionOrder <- seq_len((nrow(fem1524_sample)))
kable(fem1524_sample_id_age,
      format = 'latex',
      caption = 'Sample with Selection Order')%>%
  kable_styling(latex_options = c("hold_position"))
```

Table 3: Sample with Selection Order

ID	AGER	SelectionOrder
9	15	1
579	22	2
821	23	3
868	23	4
882	17	5
907	19	6
1145	18	7
1215	21	8
1254	22	9
1355	19	10
1718	22	11
1891	20	12
1944	24	13
2037	16	14
2283	18	15
2345	21	16
2372	20	17
2493	24	18
2638	16	19
2750	18	20

Question 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?

Code:

```
sample_estimate <- function(x, N, alpha = 0.05){
  n <- length(x)
  df <- n-1
  sample_total <- sum(x)
  sample_mean <- sum(x)/n
  sample_variance <- sum((x - rep(sample_mean, n))^2)/(n-1)
  sample_sd <- sqrt(sample_variance)
  f <- n/N
  fpc <- 1 - f
  sample_variance_est <- fpc*sample_variance/n
  sample_sd_est <- sqrt(sample_variance_est)
  tscore <- qt(1 - alpha/2, df = n-1)
  ci_lower <- sample_mean - tscore*sample_sd_est
  ci_upper <- sample_mean + tscore*sample_sd_est
  cat('Number of selected sample(n):\t\t\t', n,
      '\nDegree of Freedom:\t\t\t\t\t', df,
      '\nSample Total:\t\t\t\t\t\t', sample_total,
      '\nSample Mean:\t\t\t\t\t\t', sample_mean,
      '\nSample Variance:\t\t\t\t\t', sample_variance,
      '\nSample Standard Deviation:\t\t\t', sample_sd,
      '\nf:\t\t\t\t\t\t\t\t\t', f,
      '\nfinite population correction factor:\t', fpc,
      '\nSample Variance Estimate:\t\t\t', sample_variance_est,
      '\nSample SD Estimate:\t\t\t\t', sample_sd_est,
```

```
      '\nConfidence Interval at level ', 1-alpha,':\t [',ci_lower,',', ci_upper, ']\n'
      )
  return(list(n,
              df,
              sample_total,
              sample_mean,
              sample_variance,
              sample_sd,
              f,
              fpc,
              sample_variance_est,
              sample_sd_est,
              ci_lower,
              ci_upper))
}

fem_est <- sample_estimate(fem1524_sample$AGER, nrow(fem1524))
```

```
Number of selected sample(n):           20
Degree of Freedom:                      19
Sample Total:                           398
Sample Mean:                            19.9
Sample Variance:                        7.568421
Sample Standard Deviation:              2.751076
f:                                      0.006849315
finite population correction factor:    0.9931507
Sample Variance Estimate:               0.3758291
Sample SD Estimate:                     0.613049
Confidence Interval at level  0.95 :    [ 18.61687 , 21.18313 ]
```

Inference

For estimating the mean the minimum required statistics would be Sample Standard Deviation Estimate and Lower and Upper bound, in this case as we can see for 95% confidence interval the age is between 18.62 to 21.18 years while the point estimate being 19.9.

Question 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?

Solution

The distribution is called sampling **distribution for the sample mean.** While the standard deviation in this case is **Standard error of the Mean (SEM):**

$$SEM = \frac{\sigma}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}}$$

Question 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.

```
# Loading the data
hw1 <- read_excel('SM 625 HW 1.xlsx')

# Getting the sample of the data for the corresponding id as Q1
hw1_sample <- getdata(hw1, index)
kable(hw1_sample,
      format = 'latex',
      booktabs = TRUE,
      caption = 'Question 5 Table')%>%
  kable_styling(latex_options = c("hold_position"))
```

Table 4: Question 5 Table

ID_unit	ID	cluster	stratum	EVERPREG	AGER	PARITY	PARTS1YR
9	9	1172	117	0	15	0	1
579	579	1101	110	1	22	2	1
821	821	1311	131	0	23	0	1
868	868	1311	131	1	23	1	1
882	882	1312	131	1	17	0	1
907	907	1271	127	0	19	0	2
1145	1145	1382	138	1	18	1	4
1215	1215	1393	139	1	21	1	1
1254	1254	1401	140	1	22	1	1
1355	1355	1404	140	0	19	0	3
1718	1718	1434	143	1	22	1	1
1891	1891	1453	145	1	20	2	1
1944	1944	1461	146	1	24	3	1
2037	2037	1463	146	0	16	0	5
2283	2283	1481	148	0	18	0	1
2345	2345	1492	149	0	21	0	1
2372	2372	1493	149	1	20	2	1
2493	2493	1511	151	1	24	4	1
2638	2638	1531	153	1	16	0	3
2750	2750	1542	154	1	18	0	1

5.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, $\bar{Y}(\bar{y} = y/n = \sum_{i=1}^{20} y_i/n)$

code

```
n <- nrow(hw1_sample)
sample_mean <- sum(hw1_sample$PARTS1YR)/n
cat("The sample mean for the number of sexual male partners:\t", sample_mean)
```

```
The sample mean for the number of sexual male partners:  1.6
```

hence the population mean estimate for the number of male sexual partner are:

$$\bar{Y} = \frac{1+1+1+1+1+2+4+1+1+3+1+1+1+5+1+1+1+1+3+1}{20}$$

$$\implies \frac{32}{20} = 1.6$$

$$\bar{Y} = \bar{y} = 1.6$$

5.b

Estimate the population element variance, S^2 $[s^2 = \left[\sum_{i=1}^{20} y_i^2 - \frac{\bar{y}^2}{n}\right]\frac{1}{n-1}]$

Code:

```
sv <- sum((hw1_sample$PARTS1YR - rep(sample_mean, n))^2)/(n-1)
cat('The sample variance estimate for the number of sexual partner is:\t', sv)
```

```
The sample variance estimate for the number of sexual partner is:    1.410526
```

```
y <- data.frame(
  y = hw1_sample$PARTS1YR,
  diff = hw1_sample$PARTS1YR - sample_mean,
  squared_diff = (hw1_sample$PARTS1YR - sample_mean)^2
)
colnames(y) <- c(
  "$y$",
  "$y_i - \\bar{y}$",
  "$(y_i - \\bar{y})^2$"
)
kable(y,
      format = "latex",
      caption = "Male Sexual Partner",
      booktabs = TRUE,
      escape = FALSE) %>%
  kable_styling(latex_options = c("hold_position"))
```

hence for the 20 sample selected the estimated population variance based on the sample is 1.41

$$s^2 = \frac{.6^2 + .6^2 + .6^2 + \cdots + .6^2}{19} = \frac{.36 + .36 + \cdots + .36}{19} = 1.41$$

5.c

Estimate the sampling variance of the mean, $Var(\bar{y})$, and the standard error, $SE(\bar{y})$ $\left[var(\bar{y}) = (1-f)\frac{s^2}{n}\right.$ and $\left. se(\bar{y}) = \sqrt{var(\bar{y})}\right]$

Code:

```
f <- n/nrow(hw1)
vsem <- (1-f)*sv/n
sem <- sqrt(vsem)
cat('f:\t\t\t\t\t\t\t', f,
    '\nSample Variance of Mean:\t', vsem,
```

Table 5: Male Sexual Partner

y	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	-0.6	0.36
1	-0.6	0.36
1	-0.6	0.36
1	-0.6	0.36
1	-0.6	0.36
2	0.4	0.16
4	2.4	5.76
1	-0.6	0.36
1	-0.6	0.36
3	1.4	1.96
1	-0.6	0.36
1	-0.6	0.36
1	-0.6	0.36
5	3.4	11.56
1	-0.6	0.36
1	-0.6	0.36
1	-0.6	0.36
1	-0.6	0.36
3	1.4	1.96
1	-0.6	0.36

```
    '\nSample Standard Error:\t\t', sem)
```

```
f:                          0.006849315
Sample Variance of Mean:    0.07004326
Sample Standard Error:      0.2646569
```

the SEM is 0.264

$$SEM = \sqrt{(1 - \frac{20}{2920}) \times 1.41/20} = \sqrt{0.0988} = 0.264$$

5.d

Compute 95% confidence interval for the sample mean ($\bar{y} \pm t_{1-\alpha/2,n-1}.se(\bar{y})$)

Code:

```
tscore <- qt(1-0.05/2, n-1)
ci_lower <- sample_mean - tscore*sem
ci_upper <- sample_mean + tscore*sem
cat('The 95% confidence interval is:\t[',
    ci_lower, ',',
    ci_upper, ']')
```

```
The 95% confidence interval is: [ 1.046067 , 2.153933 ]
```

Around 95% of the number of male sexual partner falls between 1.04 to 2.15

5.e

Explain why the mean calculated in a) will not be equal to the population mean

Interpretation

1. **Sampling Variability**

- When you take a sample from a population, you are selecting only a subset of the entire population.
- The individuals included in the sample might not perfectly represent the population, leading to differences between the sample mean and the population mean.
- This variability is inherent in random sampling and is quantified by the **standard error of the mean (SEM)**.

---

2. **Finite Sample Size**

- The sample size is smaller than the population size, which increases the chance of variability in the sample mean.
- Larger samples tend to reduce this variability and produce sample means closer to the population mean, but they rarely match exactly.

---

3. **Randomness of the Sampling Process**

- The specific individuals selected in the sample are chosen randomly, and their characteristics (e.g., age, income, etc.) may differ from the population average.
- If a different random sample were drawn, the sample mean would likely differ, further illustrating this randomness.

5.f

Estimate the coefficient of variation of the mean, $CV(\bar{y})(cv(\bar{y}) = \frac{se(\bar{y})}{\bar{y}})$

```
cv <- sem/sample_mean
cat('The Coefficient of variation of the mean is CV:\t', cv)
```

```
The Coefficient of variation of the mean is CV:  0.1654105
```

$CV = \frac{0.314}{1.6} = 0.1654$

5.g

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

Interpretation

The **sampling variance of the mean** decreases as the sample size increases. This is because the **sampling variance of the mean** is inversely proportional to the sample size.

$V(\bar{y}) = \frac{\sigma^2}{n}(1 - \frac{n}{N})$ from the formula we can see that a n increases σ^2/n decreases while $1 - \frac{n}{N} \to 0$ hence as a whole

$$\lim_{n \to N} \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right) = 0$$

hence as we are increasing n from 20 to 60 we are basically decreasing the sampling variance

5.h

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

Code:

$$\begin{aligned} se(\bar{y}) = \sqrt{\frac{\sigma^2}{n}(1 - \frac{n}{N})} &= 0.05 \\ \implies \sigma^2/n - \sigma^2/N &= 0.05^2 \\ \implies 1/n - 1/N &= 0.05^2/\sigma^2 \\ \implies n &= \frac{1}{0.05^2/\sigma^2 + 1/N} \end{aligned}$$

But since $N >> n$ we can assume $1 - n/N \to 1$ hence:

$$\begin{aligned} se(\bar{y}) = \frac{\sigma}{\sqrt{n}} &= 0.05 \\ \implies n &= \frac{\sigma^2}{0.05^2} \end{aligned}$$

```
cat("The required sample size for se(y) to be 0.05 is:\t",
    round(vsem/0.05^2))
```

```
The required sample size for se(y) to be 0.05 is:    28
```

$$\begin{aligned} cv(\bar{y}) = \frac{se(\bar{y})}{\bar{y}} &= 0.1 \\ \implies se(\bar{y}) &= 0.1 \times \bar{y} \\ \implies \frac{\sigma}{\sqrt{n}} &= 0.1 \times \bar{y} \because \text{N»n} \\ \implies n &= \frac{\sigma^2}{(0.1 \times \bar{y})^2} \end{aligned}$$

```
cat('The required sample size is for cv at 0.1:\t',
    round(vsem/(0.1*sample_mean)^2))
```

```
The required sample size is for cv at 0.1:   3
```

$$
\begin{aligned}
Width &= 2 \times t \times se(\bar{y}) = 0.4 \\
&\implies 2 \times 2 \times \frac{\sigma}{\sqrt{n}} = 0.4 \because \text{N»n and give t=2} \\
&\implies (4\sigma/0.4)^2 = n
\end{aligned}
$$

```
cat("The sample size for width of 0.4 with t = 2 is:\t",
    round((4*sem/0.4)^2))
```

```
The sample size for width of 0.4 with t = 2 is:  7
```

- Sample size when $SE(\bar{y}) = 0.05$ is 28
- Sample size when $CV(\bar{y}) = 0.1$ is 3
- Sample size when width is 0.4 is 7

5.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?

Code:

```
teen <- hw1_sample %>% filter(AGER >=15 & AGER <=19) %>%
  summarize(n = n(),
            mean = mean(PARTS1YR),
            sd = sd(PARTS1YR)/sqrt(n))

kable(teen, format = 'latex', caption = 'Sexual Partner for teenagers 15-19')%>%
  kable_styling(latex_options = c("hold_position"))
```

Table 6: Sexual Partner for teenagers 15-19

n	mean	sd
9	2.333333	0.5

The mean number of sexual partners for teenagers (2.33) is higher than the overall sample mean (1.6), likely because teenagers represent a specific age group with different social and behavioral norms. Teenagers might engage in more exploratory behavior or have different relationship dynamics compared to older individuals, contributing to a higher mean. The standard error for the teenage group is larger (0.5 compared to 0.314) because the sample size for this subgroup is smaller, and there is more variability in their responses (standard error of 0.5). These differences are expected due to the unique characteristics of the teenage subgroup compared to the overall sample.