SURV625, HW-3

Sagnik Chakravarty

Table of contents

Question 1

The following are cluster totals y_α from a = 20 clusters of exactly b = 10 women (between the ages of 15 and 24) each. These clusters and the young women were sampled from the population frame used for Homework 1. The cluster totals y_α are the number of women who have ever been pregnant. Assume that the clusters were selected at random and *with replacement*, and the students were selected with *epsem* and *without replacement*. The sampling fraction is f = ab/AB = n/N = 200/2,920 = 1/14.6, meaning that the finite population correction (fpc) should not be ignored in this case.

α	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
y_α	4	4	3	6	4	6	3	4	4	1
α	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
y_α	1	8	3	3	5	6	4	5	8	5

a. Compute an estimate of the mean $\bar{y}$, its standard error, and a 95% confidence interval for the population mean. (Hint: the degrees of freedom used in computing this confidence interval should not be 199.)

Code

```
library(knitr)
library(dplyr)
library(kableExtra)
df <- data.frame(alpha = 1:20,
                 y_alpha = c(4,4,3,6,4,6,3,4,4,1,1,8,3,3,5,6,4,5,8,5))
kable(t(df), format = 'latex')
```

alpha	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y_alpha	4	4	3	6	4	6	3	4	4	1	1	8	3	3	5	6	4	5	8	5

```
y_bar <- df %>% select(y_alpha) %>% sum()/20
s2_y <- df %>%
  mutate('(y-y_bar)^2' = (y_alpha - y_bar)^2) %>%
  select(`(y-y_bar)^2`) %>% sum()/19
se <- sqrt(s2_y * (1-200/2920)/(20*100))
lbound <- y_bar - qt(p = 0.975, df = 19)*se
ubound <- y_bar + qt(p = 0.975, df = 19)*se
cat('Sample Mean:\t', y_bar,
    '\nSample Variance:\t', s2_y,
    '\nSample Standard Error:\t', se,
    '\n95% Confidence Interval:\t(', lbound, ',', ubound, ')')
```

```
Sample Mean:     4.35
Sample Variance:     3.502632
Sample Standard Error:   0.04039013
95% Confidence Interval:    ( 4.265462 , 4.434538 )
```

Calculation

$$\text{Sample Mean: } \bar{y} = \frac{1}{\alpha}\sum_{\alpha=1}^{20} y_\alpha = \frac{87}{20} = 4.35$$

$$\text{Sample Variance: } s_y^2 = \frac{1}{\alpha-1}\sum_{\alpha=1}^{20}(y_\alpha - \bar{y}) = \frac{1}{19}\sum(y_\alpha - 4.35) = \frac{66.55}{19} \approx 3.503$$

$$\text{Standard Error: } SE(\bar{y}) = \sqrt{\frac{s_y^2}{\alpha b^2} \times (1-f)} = \sqrt{\frac{3.503}{20 \times 10^2} \times \left(1 - \frac{20}{2920}\right)} \approx 0.0404$$

$$\text{95\% Confidence Interval: } CI = \bar{y} \pm t_{1-\alpha/2,df} \times SE(\bar{y}) = 4.35 \pm 2.093 \times 0.404 = [3.505, 5.195]$$

The sample mean $\bar{y}$ of women who have ever been pregnant across 20 clusters is **4.35** with a standard error (SE) of **0.0404**. The 95% confidence interval (CI) for the population mean is [**4.265462 , 4.434538**]. This interval reflects uncertainty accounting for the cluster sampling design, where degrees of freedom (df = 19) align with the number of clusters rather than individual observations1. The finite population correction (fpc) factor of 0.931 slightly narrows the CI compared to a design without fpc.

b. Estimate the standard error of the mean that you would expect if the sample consisted of a = 40 clusters of b = 10 each. (Hint: What about this design has not changed, and what quantity needed to answer this question could therefore be considered portable?)

Code

```
se_new <- sqrt((1-400/2920)*s2_y/(40*10^2))
cat('The new standard error with a = 40 is:\t', se_new)
```

```
The new standard error with a = 40 is:   0.02749008
```

Calculation

$$SE_{new}(\bar{y}) = \sqrt{\frac{(1-f_{new})s_y^2}{\alpha_{new} \times b^2}} = \sqrt{\frac{(1-\frac{400}{2920})3.503}{40 \times 10^2}} = 0.0275$$

Increasing the number of clusters from 20 to 40 (while maintaining 10 women per cluster) reduces the standard error to 0.0275, a 32% decrease from the original SE. This improvement stems from the inverse relationship between cluster count and variance:

$$SE \propto \frac{1}{a}$$

where a is the number of clusters. The portable quantity here is the between-cluster variance:

$$s_y^2 = 3.503$$

which remains stable under the assumption of similar cluster homogeneity.

c. Note that the mean $\bar{y}$ is a proportion. Based on the sample of 20 clusters [and ignoring the ratio n / (n – 1)], compute the design effect deff, as well as roh. How would you interpret the design effect for a colleague in plain English?

Code

```
p <- sum(df$y_alpha)/200
var_srs <- p*(1-p)*(1-200/2920)/200
var_cluster <- se^2
deff <- var_cluster/var_srs
roh <- (deff-1)/9

cat('Design Effect:\t', deff,
    '\nRate of Heterogenity:\t', roh)
```

```
Design Effect:   1.425137
Rate of Heterogenity:    0.04723749
```

Calculation

$$\text{p} = \frac{\text{Total Pregnencies in Sample}}{\text{Total Women Sampled}} = \frac{87}{200} = 0.435$$

$$var_{srs} = \frac{p(1-p)}{n}(1-f) = \frac{0.435 \times 0.565}{20} \times (1 - \frac{200}{2920}) = 0.001145$$

$$var_{cluster} = \frac{s_y^2}{\alpha b^2} \times (1-f) = \frac{3.503}{20 \times 10^2} \times \left(1 - \frac{20}{2920}\right) = 0.00163$$

Design Effect

$$deff = \frac{var_{cluster}}{var_{srs}} = \frac{0.001145}{0.00163} = 1.425$$

Rate of Heterogenity

$$rho = \frac{deff - 1}{b - 1} = \frac{1.425 - 1}{10 - 1} = \frac{0.425}{9} = 0.0472$$

The design effect ($deff = 1.425$) indicates that cluster sampling introduces 42.5% more variance compared to simple random sampling (SRS). This inefficiency arises from similarities within clusters, quantified by the intraclass correlation coefficient ($\rho = 0.047$).

Interpreting ρ, approximately 4.7% of the total variance in pregnancy status is attributable to between-cluster differences. For practical survey design, this implies that while clustering introduces measurable inefficiency, the effect is moderate.

d. Now, using the computed value of roh from part (c), estimate the standard error that you would expect from a sample of a = 40 clusters of b = 5 women each.

Code

```
deff_new <- 1 + (5-1)*roh
f_new <- 40*5/2920
var_new <- deff_new*var_srs
cat('The new stadard error is:\t', sqrt(var_new))
```

```
The new stadard error is:    0.0368917
```

Calculation

$$
\begin{aligned}
&\text{New Design Effect} \\
&deff_{new} = 1 + (b_{new} - 1) \cdot roh = 1 + (5 - 1) \cdot 0.0472 = 1.18895 \\
\\
&\text{Adjusted Variance} \\
&var_{new} = deff_{new} \times var_{srs} = 1.189 \times 0.001145 = 0.00136 \\
\\
&\text{New Standard Error} \\
&SE_{new} = \sqrt{var_{new}} = \sqrt{0.00136} = 0.0369
\end{aligned}
$$

For a modified design with 40 clusters of 5 women each, the standard error increases to 0.0369 due to reduced cluster size. This result leverages the previously calculated ρ to estimate the new design effect ($deff_{new} = 1.189$).

The trade-off between cluster size and count highlights the importance of optimizing survey designs to balance cost and precision.

Question 2

The data set, cherry.csv, contains measurements of diameter (inches), height (feet), and timber volume (cubic feet) for a sample of 31 black cherry trees. Diameter and height of trees are easily measured, but volume is more difficult to measure.

Data

```r
cherry <- read.csv('cherry.csv')
kable(t(cherry), format = 'latex', booktabs = TRUE) %>%
  kable_styling(latex_options = c("hold_position", "scale_down"))
```

diameter	8.3	8.6	8.8	10.5	10.7	10.8	11.0	11.0	11.1	11.2	11.3	11.4	11.4	11.7	12.0
height	70.0	65.0	63.0	72.0	81.0	83.0	66.0	75.0	80.0	75.0	79.0	76.0	76.0	69.0	75.0
volume	10.3	10.3	10.2	16.4	18.8	19.7	15.6	18.2	22.6	19.9	24.2	21.0	21.4	21.3	19.1

diameter	12.9	12.9	13.3	13.7	13.8	14.0	14.2	14.5	16.0	16.3	17.3	17.5	17.9	18.0	18	20.6
height	74.0	85.0	86.0	71.0	64.0	78.0	80.0	74.0	72.0	77.0	81.0	82.0	80.0	80.0	80	87.0
volume	22.2	33.8	27.4	25.7	24.9	34.5	31.7	36.3	38.3	42.6	55.4	55.7	58.3	51.5	51	77.0

```r
print(dim(cherry))
```

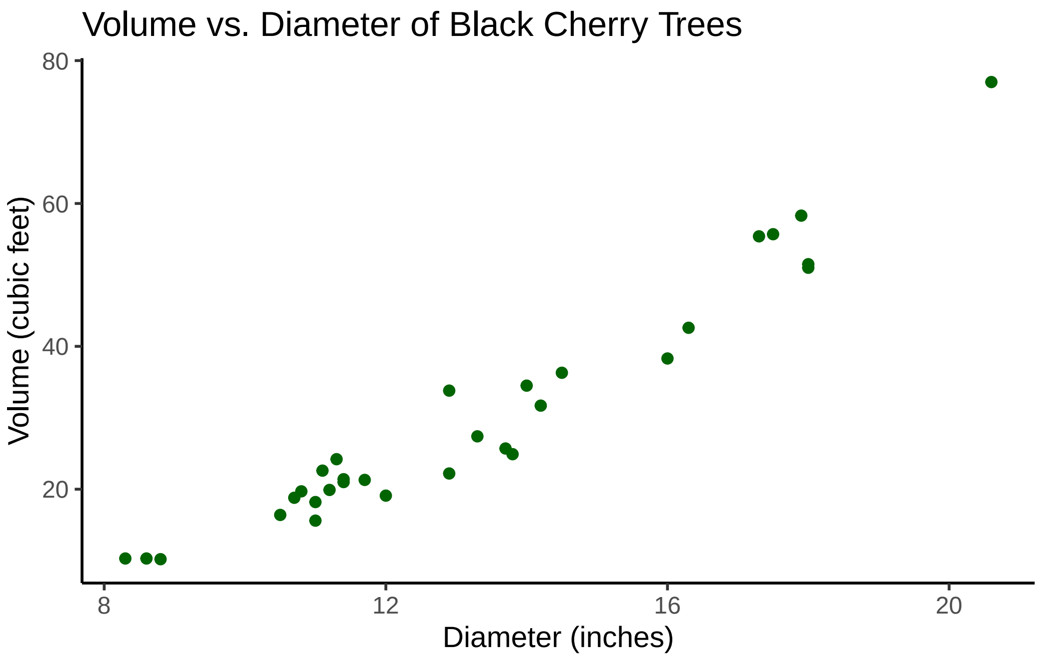
```
[1] 31  3
```

a. Plot volume vs. diameter for the 31 trees.

Code

```r
library(ggplot2)
cherry <- read.csv("cherry.csv")
ggplot(cherry, aes(x = diameter, y = volume)) +
  geom_point(color = "darkgreen") +
  labs(
    title = "Volume vs. Diameter of Black Cherry Trees",
    x = "Diameter (inches)",
    y = "Volume (cubic feet)"
  ) +
```

```
    theme_classic()
```



Interpretation

The plot shows a strong positive linear relationship between tree diameter and timber volume. Larger diameters correlate with higher volumes, justifying the use of ratio/regression estimation.

b. Suppose that these trees are an SRS from a forest of N = 2967 trees and that the sum of the diameters for all trees in the forest is 41,835 inches. Use ratio estimation to estimate the total volume for all trees in the forest. Give a 95% CI.

Code

```r
n <- nrow(cherry)
N <- 2967
T_x <- 41835
B_ratio <- sum(cherry$volume) / sum(cherry$diameter)
t_ratio <- B_ratio * T_x
residuals_ratio <- cherry$volume - B_ratio * cherry$diameter
s_r_sq <- sum(residuals_ratio^2) / (n - 1)
se_ratio <- N * sqrt((1 - n/N) * s_r_sq / n)
t_value_ratio <- qt(0.975, df = n - 1)
CI_ratio <- t_ratio + c(-1, 1) * t_value_ratio * se_ratio

cat('Ratio Estimator:\t', t_ratio,
    '\nStandard Error:\t', se_ratio,
    '\n95% Confidence Interval:\t', CI_ratio)
```

```
Ratio Estimator:     95272.16
Standard Error:  5140.933
95% Confidence Interval:     84772.97 105771.3
```

Calculation

$$\hat{t}_y = \frac{\sum y_i}{\sum x_i} \cdot T_x = 95272.16$$
$$SE(\hat{t}_y) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n}} = 5140.933 \text{ where } s_r^2 = \frac{1}{n-1}\sum(y_i - Bx_i)^2$$
$$CI = \hat{t}_y \pm t_{1-\alpha/2,n-2} SE(\hat{t}_y) = (84772.97, 105771.3)$$

c. Use regression estimation to estimate the total volume for all trees in the forest. Give a 95% CI.

Code

```
model <- lm(volume ~ diameter, data = cherry)
beta <- coef(model)[2]
y_bar <- mean(cherry$volume)
x_bar_sample <- mean(cherry$diameter)
X_bar_pop <- T_x / N
t_reg <- N * (y_bar + beta * (X_bar_pop - x_bar_sample))
mse <- sum(residuals(model)^2) / (n - 2)
se_reg <- N * sqrt((1 - n/N) * mse / n)
t_value_reg <- qt(0.975, df = n - 2)
CI_reg <- t_reg + c(-1, 1) * t_value_reg * se_reg

cat('The regression estimate being:\t', t_reg,
    '\n95% CI:\t', CI_reg)
```

```
The regression estimate being:   102318.9
95% CI:  97708.98 106928.7
```

Calculations

$$\hat{t}_y = N\bar{y} + \hat{\beta}(T_x - N\bar{x}) = 102318.9$$
$$SE(\hat{t}_y) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{MSE}{n}} = 2253.97$$
$$CI = \hat{t}_y \pm t_{1-\alpha/2,n-2} SE(\hat{t}_y) = (97708.98, 106928.7)$$