SURV625, HW-2

Sagnik Chakravarty

Table of contents

Question 1

A stratified random sample of graduates from the academic departments at a small university is to be selected to estimate the mean starting salary of graduates from that university. The university has five academic departments, and preliminary data are obtained from an attempted census in a recent year:

Stratum	Department	N_h	$\bar{Y}_h$	S_h
1	Humanities	20	46,800	81,200
2	Social Sciences	90	61,500	101,700
3	Natural Sciences	120	76,100	130,900
4	Engineering	140	89,500	170,100
5	Business	200	95,500	216,400

a. What is the mean starting salary $\bar{Y}$ per graduate across all departments in the population, based on the data collected from this previous attempted census?

Solution

```r
library(knitr)
library(dplyr)
library(kableExtra)
df <- data.frame(stratum = 1:5
                 ,Department = c('Humanities', 'Social Sciences', 'Natural Sciences', 'Engineer
                 N_h = c(20, 90, 120, 140, 200),
                 Y_bar_h = c(46800, 61500, 76100, 89500, 95500),
                 S_h = c(81200, 101700, 130900, 170100, 216400))
df['W_h'] <- df$N_h/sum(df$N_h)
df['W_h.Y_h'] = df$W_h*df$Y_bar_h
kable(df, format = 'latex')
```

stratum	Department	N_h	Y_bar_h	S_h	W_h	W_h.Y_h
1	Humanities	20	46800	81200	0.0350877	1642.105
2	Social Sciences	90	61500	101700	0.1578947	9710.526
3	Natural Sciences	120	76100	130900	0.2105263	16021.053
4	Engineering	140	89500	170100	0.2456140	21982.456
5	Business	200	95500	216400	0.3508772	33508.772

```r
cat('the mean starting salary accross all departments:\t', sum(df$W_h.Y_h))
```

```
the mean starting salary accross all departments:    82864.91
```

Calculation

The mean starting salary per graduate across all departments in the population is the weighted mean of the stratum means, given by:

$$\bar{Y} = \sum W_h \bar{Y}_h$$

where:

- $W_h = \frac{N_h}{N}$ is the proportion of graduates in stratum h
- N_h is the number of graduates in department h

- $\bar{Y}_h$ is the mean starting salary for department h
- $N = \sum_h N_h$ is the total number of graduates.

$$
\begin{aligned}
N &= 20 + 90 + 120 + 140 + 200 = 570 \\
W_h &= \frac{20}{570}, \frac{90}{570}, \frac{120}{570}, \frac{140}{570}, \frac{200}{570} \\
\bar{Y} &= \frac{20}{570} \times 46,800 + \frac{90}{570} \times 61,500 + \frac{120}{570} \times 76,100 + \frac{140}{570} \times 89,500 + \frac{200}{570} \times 95,500 = 82864.91
\end{aligned}
$$

Hence $\bar{Y} = 82,864.91$

b. What is the average within-stratum element variance $S_w^2 = \sum W_h S_h^2$

Solution

```
df['S_h^2'] = df$S_h^2
cat('The average within stratum element variance:\t',
    sum(df$`S_h^2`*df$W_h))
```

```
The average within stratum element variance:     29009578070
```

Calculation

$$
\begin{aligned}
S_h^2 &= \frac{20}{570} \times 81,200^2 + \frac{90}{570} \times 101,700^2 + \frac{120}{570} \times 130,900^2 + \frac{140}{570} \times 170,100^2 + \frac{200}{570} \times 216,400^2 \\
&= 29009578070
\end{aligned}
$$

hence 2.9×10^{10}

c. For a sample of $n = 100$, what is the proportionate allocation?

Solution

```
df['proportionate'] <- round(df$W_h *100)
kable(df[c('stratum', 'Department', 'proportionate')], format = 'latex')
```

stratum	Department	proportionate
1	Humanities	4
2	Social Sciences	16
3	Natural Sciences	21
4	Engineering	25
5	Business	35

Calculation

In proportionate allocation, the sample size for each stratum n_h is determined based on the proportion of the population in that stratum:

$$n_h = W_h n$$

where

- $W_h = \frac{N_h}{N}$
- n is the number of sample according to the question $n = 100$

$$\begin{aligned} n_1 &= \frac{20}{570} \times 100 = 3.51 \approx 4 \\ n_2 &= \frac{90}{570} \times 100 = 15.79 \approx 16 \\ n_3 &= \frac{120}{570} \times 100 = 21.05 \approx 21 \\ n_4 &= \frac{140}{570} \times 100 = 24.56 \approx 25 \\ n_5 &= \frac{200}{570} \times 100 = 35.09 \approx 35 \end{aligned}$$

d. For a sample of $n = 100$, what is the Neyman allocation?

Solution

```
df['W_hS_h'] = df$W_h*df$S_h
df['Neyman'] = round(100 * df$W_hS_h/sum(df$W_hS_h))
kable(df[c('stratum', 'Department', 'Neyman')], format = 'latex')
```

stratum	Department	Neyman
1	Humanities	2
2	Social Sciences	10
3	Natural Sciences	17
4	Engineering	25
5	Business	46

Calculation

$$n_h = k.W_h S_h = \frac{W_h S_h}{\sum_h W_h S_h} \times n \text{ where } k = \frac{n}{\sum_h W_h S_h}$$

Stratum	Department	$W_h S_h$	Calculation	n_h
1	Humanities	$\frac{20}{570} \times 81,200 = 2849.123$	$\frac{2849.123}{164173.7} \times 100$	2
2	Social Sciences	$\frac{20}{570} \times 101,700 = 16,057.895$	$\frac{16,057.895}{164173.7} \times 100$	10
3	Natural Sciences	$\frac{20}{570} \times 130,900 = 27,557.895$	$\frac{27,557.895}{164173.7} \times 100$	17
4	Engineering	$\frac{20}{570} \times 170,100 = 41,778.947$	$\frac{41,778.947}{164173.7} \times 100$	25
5	Business	$\frac{20}{570} \times 216,400 = 75,929.825$	$\frac{75,929.825}{164173.7} \times 100$	46

Proportionate Variance	Neyman Variance
237958239	218775808

e. Estimate the sampling variance of the mean for the proportionate c) and Neyman d) allocations.

Solution

```
kable(df %>%
  mutate('prop_variance' = W_h^2*(1-proportionate/N_h)*`S_h^2`/proportionate,
         'neyman_variance' = W_h^2*(1-Neyman/N_h)*`S_h^2`/Neyman) %>%
  summarize('Proportionate Variance' = sum(prop_variance),
            'Neyman Variance' = sum(neyman_variance)),
  format = 'latex', booktabs = TRUE) %>%
  kable_styling(latex_options = "scale_down")
```

Calculation

$$var(\bar{y}_w) = \sum_h W_h^2 \frac{1-f}{n_h} s_h^2 \text{ where } f = \frac{n_h}{N_h}$$

Stratum	Department	Proportionate Variance	Neyman Variance
1	Humanities	$(\frac{20}{570})^2(1-\frac{4}{20})\frac{81200^2}{4} = 1,623,500$	$(\frac{20}{570})^2(1-\frac{2}{20})\frac{81200^2}{2} = 3,652,875$
2	Social Sciences	13,250,932	22,920,532
3	Natural Sciences	29,835,047	38,344,151
4	Engineering	57,351,500	57,351,500
5	Business	135,897,259	96,506,749

hence

- **Proportionate Variance** = 237958239
- **Neyman Variance** = 218775808

f. Estimate the total element variance S^2

Solution

```
df %>%
  mutate('S2' = (N_h-1)/(sum(N_h) - 1)*(`S_h^2` + (Y_bar_h - 82864.91)^2)) %>%
  summarise('S^2' = sum(S2)) %>% kable(format = 'latex')
```

S^2
29058520065

Calculation

$$S^2 = \sum_h \frac{N_h - 1}{N-1} S_h^2 + \sum_h \frac{N_h - 1}{N-1}(\bar{Y}_h - \bar{Y})^2 = 29058520065$$

hence $S^2 = 2.9 \times 10^{10}$

g. What are the design effects of the proportionate and Neyman allocations?

Solution

```
cat('Design Effect of Proportion Allocation:\t', 237958239/2905852006,
    '\nDesign Effect of Neyman Allocation:\t', 218775808/2905852006)
```

```
Design Effect of Proportion Allocation:  0.08188932
Design Effect of Neyman Allocation:  0.07528801
```

Calculation

$$deff = \frac{var(\bar{y})}{var_{SRS}(\bar{y})} = \frac{s_w^2}{s^2}$$

hence

- Design Effect of Proportion Allocation: **0.08188932**
- Design Effect of Neyman Allocation: **0.07528801**

h. Suppose that the cost-per-element was not the same in each stratum: $C_1 = C_2 = C_3 = \$30, C_4 = C_5 = \$40$

1. The client requesting a stratified sample design has indicated that the total available data collection budget is = \$5,000, with the stratum specific costs per element listed above. What allocation will minimize the sampling variance of the mean under these cost constraints?

Solution

```
df['Cost'] <- c(30, 30, 30, 40, 40)
k_cost <- 5000/sum((df$W_h*df$S_h*sqrt(df$Cost)))
df <- df %>%
  mutate('Cost Allocation' = round(k_cost*W_h*S_h/sqrt(Cost)))
df %>%
  select(stratum, Department, Cost, `Cost Allocation`) %>%
  kable(format = 'latex')
```

stratum	Department	Cost	Cost Allocation
1	Humanities	30	3
2	Social Sciences	30	15
3	Natural Sciences	30	25
4	Engineering	40	33
5	Business	40	60

Calculation

$$C = 5000$$
$$k = \frac{C}{\sum_h W_h S_h \sqrt{C_h}} = \frac{5000}{0.035 \times 81200 \times \sqrt{30} + 0.157 \times 101700 \times \sqrt{30} + \cdots} = 0.005005233$$
$$n_h = \frac{kW_hS_h}{\sqrt{C_h}} = (\frac{0.005 \times 0.035 \times 81200}{\sqrt{0.005}}, \cdots) = (3, 15, 25, 33, 60)$$

2. Estimate the expected sampling variance and design effect of the mean starting salary under this allocation

Solution

```
df %>%
  mutate('cost_variance' = W_h^2*(1-`Cost Allocation`/N_h)*`S_h^2`/`Cost Allocation`) %>%
  summarize('Cost Allocation Variance' = sum(cost_variance),
            'design effect' = sum(cost_variance)/2905852006) %>%
  kable(format = 'latex')
```

Cost Allocation Variance	design effect
148362056	0.0510563

Calculation

$$var(\bar{y}_w) = \sum_h W_h^2 \frac{1-f}{n_h} s_h^2 = 148362056$$

$$\text{Design Effect} = \frac{var(\bar{y})}{var_{SRS}(\bar{y})} = \frac{s_w^2}{s^2} = \frac{148362056}{2905852006} = 0.051$$