BST5420 Sampling Theory and Survey Design Homework 2

Miao Cai[*]

2019-03-05

Note that the points total 30, but the assignment is worth 10 points. Your homework score will be recorded as your score out of 30, divided by 3.

1. (11 Points: 2,3,3,3) Researchers are interested in the proportion of drivers in the city who use their cell phone by holding it with their hands while driving (i.e., not using a blue-tooth, or hands-free device). Investigators chose a busy intersection, and observed every sixth car that passed going a particular direction. It was suggested that they look at every car and record whether the driver was using the cell phone, but often cars went by too fast and they could misidentify some drivers. So, the plan was to dispatch two investigators, one of whom counted the cars and chose every sixth car and the other looked carefully to see whether the driver was talking on the cell phone. They did this for two hours and observed 300 drivers.

   1. What kind of a sample is this?
   2. Is it a reasonable sampling plan for determining the rate of cell phone usage while driving at that intersection at that time?
   3. Is this a reasonable plan for determining the rate of cell phone usage while driving for the city as a whole?
   4. Can you describe a better sampling plan? Explain.

Answers:

1. This is systematic sampling since it chooses every 6th car in the intersection.
2. Since I do not see a clear periodic pattern of car flow at the intersection, I think this is a reasonable plan.
3. No, this is not a good plan for the city as a whole. This is a busy intersection and people are less likely to use cell phone while driving. The rate of cell phone usage while driving at this intersection will tend to underestimate the rate since it does not account for less busy intersections, where people are more likely to use cell phone while driving.
4. A better sampling plan could be using a two-statge plan:

- In the first stage, take a random sample of intersections at different hours of the day.
- In the second stage, do the systematic sample as described in the question.

The reason why this plan is better is that it randomly choose intersections (clusters) at different hours in the city first, this accounts for the huge difference of traffic flows at different intersections and different hours (rush hours and non-rush hours).

[*] PhD student, Department of Epidemiology and Biostatistics, College for Public Health and Social Justice, Saint Louis University. Email address miao.cai@slu.edu

2. (9 Points: 3,3,3) Do Problem 4.3 in the textbook. This problem begins “Foresters want to estimate . . . " only parts (a) and (b).

Foresters want to estimate the average age of trees in a stand. Determining age is cumbersome, because one needs to count the tree rings on a core taken from the tree. In general, though, the older the tree, the larger the diameter, and diameter is easy to measure. The foresters measure the diameter of all 1132 trees and find that the population mean equals 10.3. They then randomly select 20 trees for age measurement.

1. Draw a scatterplot of y vs. x.
2. Estimate the population mean age of trees in the stand using ratio estimation and give an approximate standard error for your estimate.
3. Repeat (b) using regression estimation.
4. Label your estimates on your graph. How do they compare?

Answers:

1. R code:

```
pacman::p_load(tidyverse)
options(digits = 3, scipen = 999)

Diameter = c(
  12, 11.4, 7.9, 9, 10.5, 7.9, 7.3, 10.2, 11.7, 11.3,
  5.7, 8.0, 10.3, 12, 9.2, 8.5, 7, 10.7, 9.3, 8.2)
Age = c(125, 119, 83, 85, 99, 117, 69, 133, 154, 168,
        61, 80, 114, 147, 122, 106, 82, 88, 97, 99)

dat = data.frame(Diameter, Age)

ggplot(dat, aes(Diameter, Age)) +
  geom_point() + theme_bw()
```

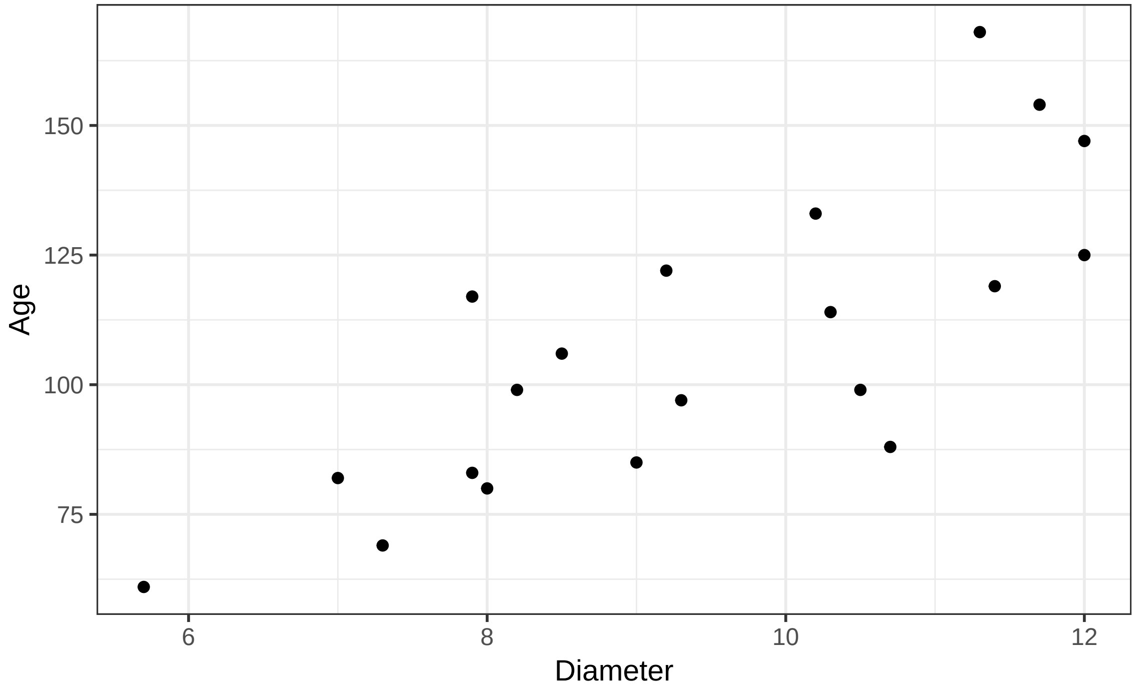

Figure 1: Scatterplot of age over diameter in the 20 sample trees

2. Among the 20 sample trees:

- $\bar{y} = 107.4$,
- $\bar{x} = 9.405$
- $\hat{B} = \hat{\tau}_y / \hat{\tau}_x = \bar{y}/\bar{x} = 11.419$

The population mean age is then: - $\hat{\mu}_y = \hat{B}\mu_x = 117.62$

The standard error of the ratio estimate can be calculated using the following formula and R code:

$$e_i = y_i - \hat{B}x_i$$

$$s_e^2 = \frac{1}{n-1} \sum_{i \in S} e_i^2$$

$$\widehat{V(\hat{\tau})} = (1 - \frac{n}{N}) * \tau_x^2 \frac{S_e^2}{n\bar{x}^2}$$

R code:

```r
N = 1132
n = length(Age)
tau_x = 10.3

Bhat = mean(Age)/mean(Diameter)
residual = Age - Bhat*Diameter
SSE = sum(residual^2)/(n-1)
VhatB = (1-n/N)*SSE*tau_x^2/(n*mean(Diameter)^2)
standard_error = sqrt(VhatB)
```

The standard error of the mean age $\widehat{sd(\hat{\tau})}$ is: 4.355

3. (10 Points) Do Problem 5.7 in the textbook. The problem begins “The new candy Green . . . ”. Be sure to estimate the total number of cases and the average number sold per store.

The new candy Green Globules is being test-marketed in an area of upstate NewYork. The market research firm decided to sample 6 cities from the 45 cities in the area and then to sample supermarkets within cities, wanting to know the number of cases of Green Globules sold.

City	Number of Supermarkets	Number of Cases Sold
1	52	146, 180, 251, 152, 72, 181, 171, 361, 73, 186
2	19	99, 101, 52, 121
3	37	199, 179, 98, 63, 126, 87, 62
4	39	226, 127, 57, 46, 86, 43, 85, 165
5	8	12, 23
6	14	87, 43, 59

Obtain summary statistics for each cluster. Plot the data, and estimate the total number of cases sold, and the average number sold per supermarket, along with the standard errors of your estimates.

1. Summary statistics for each cluster:

```r
pacman::p_load(tidyverse)

city = c(rep(1, 10), rep(2, 4), rep(3, 7),
         rep(4, 8), rep(5, 2), rep(6, 3))
CaseNumber = c(
  146, 180, 251, 152, 72, 181, 171, 361, 73, 186,
  99, 101, 52, 121,
  199, 179, 98, 63, 126, 87, 62,
  226, 129, 57, 46, 86, 43, 85, 165,
  12, 23,
  87, 43, 59
)

dat3 = data.frame(city, CaseNumber)
dat3$city = as.character(dat3$city)

sum_tab = dat3 %>%
  group_by(city) %>%
  summarise(
    m_i = n(),
    Mean = mean(CaseNumber),
    Std_dev = sd(CaseNumber),
    Sum = sum(CaseNumber))
```

```r
knitr::kable(
  sum_tab, align = 'c',
  caption = 'Summary statistics for each cluster',
  digits = 2)
```

Table 1: Summary statistics for each cluster

city	m_i	Mean	Std_dev	Sum
1	10	177.3	83.60	1773
2	4	93.2	29.24	373
3	7	116.3	54.54	814
4	8	104.6	64.59	837
5	2	17.5	7.78	35
6	3	63.0	22.27	189

Boxplots for each cluster:

```r
ggplot(dat3, aes(x = city, y = CaseNumber)) +
  geom_boxplot(color = "#1F3552", size = 0.9) +  theme_bw() +
  ylab("Total number of candy Green Globules sold")
```

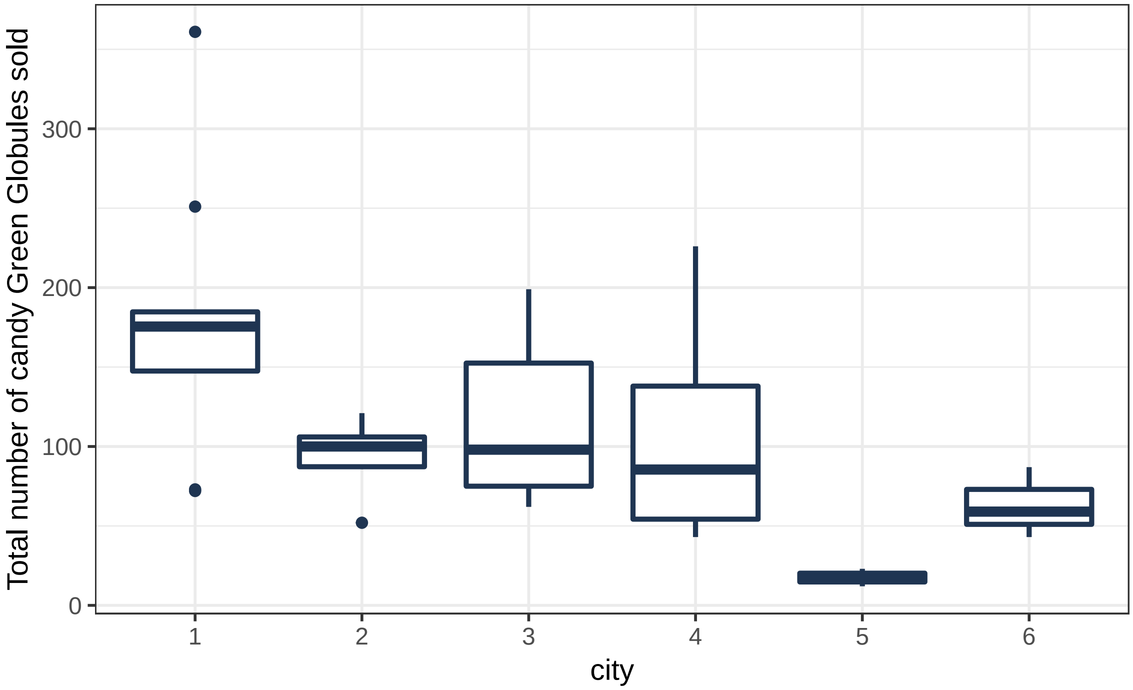

Figure 2: Boxplots of the total number of candy Green Globules sold in each sample city

In this case, the total number of elements M_i in the population is unknown. The M can be estimated using the following formula:

$$\hat{M} = \frac{N}{n} \sum_{i \in S} M_i$$

Then $\hat{\tau}$ and $\hat{\mu}$ can be estimated using the following formulas and R code:

$$\hat{\tau} = \frac{N}{n}\sum_{i\in S} M_i\bar{y}_i$$

$$\hat{\mu} = \frac{\hat{\tau}}{M} = \frac{\hat{\tau}}{M}\frac{N}{n}\sum_{i\in S} M_i\bar{y}_i$$

```r
M_i = c(52, 19, 37, 39, 8, 14)
m_i = sum_tab$m_i
N = 45
n = nrow(sum_tab)

M_hat = N/n*sum(M_i)
y_bar = sum_tab$Mean
tau_hat = N/n*sum(M_i*y_bar)
mu_hat = N/n*sum(M_i*y_bar)/M_hat
```

The estimated values are:

- $\hat{\tau} = 152972.223$
- $\hat{\mu} = 120.688$

To estimate the standard errors of the two estimates, we need to calculate three intermediate variables:

$$\hat{\tau}_i = \frac{M_i}{m_i}\sum_{j\in S_i} y_{ij}$$

$$s_t^2 = \frac{1}{n-1}\sum_{i\in S}(\hat{\tau}_i - \frac{\hat{\tau}}{N})^2$$

$$s_i^2 = \frac{1}{m_i-1}\sum_{j\in S_i}(y_{ij} - \bar{y}_i)^2$$

Then, $\widehat{V(\hat{\tau})}$ and $\widehat{V(\hat{\mu})}$ can be estimated using the following formula:

$$\widehat{V(\hat{\tau})} = N^2(1-\frac{n}{N})\frac{S_t^2}{n} + \frac{N}{N}\sum_{i\in S}(1-\frac{m_i}{M_i}M_i^2\frac{s_i^2}{m_i})$$

$$\widehat{V(\hat{\mu})} = \frac{N^2}{M^2}(1-\frac{n}{N})\frac{s_r^2}{n} + \frac{N}{M^2 n}\sum_{i\in S}(1-\frac{m_i}{M_i})M_i^2\frac{s_i^2}{m_i}$$

```r
tau_i = M_i/m_i*sum_tab$Sum
s_t_square = 1/(n - 1)*sum((tau_i - tau_hat/N)^2)

sum_square = sum_tab %>%
  select(city, Mean) %>%
  right_join(dat3) %>%
  mutate(diff_square = (CaseNumber - Mean)^2) %>%
  group_by(city) %>%
  summarise(sum_square = sum(diff_square)) %>%
  pull(sum_square)
```

```
## Joining, by = "city"
```

```
s_i_square = 1/(m_i - 1)*sum_square
V_tau = N^2*(1 - n/N)*s_t_square/n + N/n*sum((1 - m_i/M_i)*M_i^2*s_i_square/m_i)
sd_tau = sqrt(V_tau)

# V_mu
s_r_square = sum((M_i*y_bar - M_i*mu_hat)^2)
V_mu = N^2/M_hat^2*(1 - n/N)*s_r_square/n +
  N/(M_hat^2*n)*sum((1 - m_i/M_i)*M_i^2*s_i_square/m_i)
sd_mu = sqrt(V_mu)
```

The estimates of the standard errors are:

- $\widehat{V(\hat{\tau})} = 3224091076.202$, $\widehat{sd(\hat{\tau})} = 56781.08$,
- $\widehat{V(\hat{\mu})} = 1959.073$, $\widehat{sd(\hat{\mu})} = 44.261$.