STA304 A3

Yu-Chun Chien

4/3/2021

Question 1

(a) Estimate of Average Repair Cost Per Saw

Estimate of average repair cost per saw for the past month with a bound on the error of estimation: $\bar{y} \pm 2\sqrt{\hat{Var}(\bar{y})} = 19.73077 \pm 2 \times 0.8900517 = (17.95067, 21.51087)$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = \frac{2565}{130} = 19.73077$$

$$s_r^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}m_i)^2}{n-1} = \frac{16065.65}{19} = 845.5607$$
$$\hat{Var}(\bar{y}) = (1 - \frac{n}{N})\frac{s_r^2}{n\bar{m}^2} = (1 - \frac{20}{96})\frac{845.607}{20 \times 6.5^2} = 0.792192$$

(b) Number of Clusters Needed If Bound Is Less Than \$2

$n = \frac{N\sigma_r^2}{ND + \sigma_r^2}$, where $D = \frac{B^2\bar{M}^2}{4}$

Here σ_r^2 is estimated by s_r^2, $\bar{M}$ is estimated by $\bar{m}$ and $D = \frac{2^2 \times 6.5^2}{4} = 42.25$, so $n = 16.56$

Thus, for B to be less than \$2, 17 clusters should be selected.

(c) Relationship Between n and B

In part (a), $B = 1.6$ and $n = 20$

In part (b), $B = 2$ and $n = 17$

The number of clusters in part (b) is smaller and the bound is wider. Thus, as n decreases, B increases. In other words, we need more clusters in order to have a more precise estimation.

Question 2

(a) Estimate of Average Sales for All Supermarkets

Estimate of average sales for the week for all supermarkets in the area with a bound on the error of estimation: $\bar{y} \pm 2\sqrt{\hat{Var}(\bar{y})} = 97.9728 \pm 2 \times 5.497766 = (86.97727, 108.9683)$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} M_i\bar{y}_i}{\sum_{i=1}^{n} M_i} = \frac{14402}{147} = 97.9728$$

$s_r^2 = \frac{\sum_{i=1}^{n} M_i^2(\bar{y}_i-\hat{\mu})^2}{n-1} = \frac{693853.3}{4} = 173463.3$, while s_i^2 is given
$\hat{V}ar(\hat{\mu}) = (1-\frac{n}{N})\frac{s_r^2}{n\bar{M}^2} + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n} M_i^2(1-\frac{mi}{Mi})(\frac{s_i^2}{m_i}) = 30.22544$

The estimator is not unbiased. Since we do not know M (population size), we need to estiate M from the sample data. In this case, we are using ratio estimation, which is biased. (obtain $\bar{M}$ by $\frac{\sum_{i=1}^{n} M_i}{n}$)

(b) Estimate of Total Number of Boxes Sold

Yes, estimate it by $\hat{\tau} = M\hat{\mu} = \frac{N}{n}\sum_{i=1}^{n} M_i\bar{y}_i$

$\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n} M_i\bar{y}_i = 57608$

$\hat{V}ar(\hat{\tau}) = N^2 \times \bar{M}^2 \times \hat{V}ar(\bar{y}) = 10450252.309225$

Estimate of total sales for the week for all supermarkets in the area with a bound on the error of estimation: $\hat{\tau} \pm 2\sqrt{\hat{V}ar(\hat{\tau})} = 57608 \pm 2 \times 3232.685 = (51142.63, 64073.37)$

Question 3

(a) Simple Random Sample

```
set.seed(1005194380)
sample_yc <- data_yc[sample(210,10),]
sample_yc
```

```
##      id height handspan
## 47   49  160.0    20.00
## 126 134  169.5    20.55
## 134 142  168.0    18.50
## 196 208  152.4    21.00
## 19   20  178.0    24.00
## 106 113  160.0    19.20
## 30   31  162.0    18.20
## 66   70  175.0    21.00
## 67   71  168.0    15.00
## 121 128  168.0    19.00
```

(b) SRS Estimate of Mean

$\mu_y = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = 167.54$

$\hat{V}ar(\bar{y}) = (1-\frac{n}{N})\frac{s^2}{n} = 7.9179$, where $s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = 83.13822$

Estimate of μ_y with bound on error of estimation is $\bar{y} \pm 2\sqrt{\hat{V}ar(\bar{y})} = 167.54 \pm 2 \times 2.8139 = (161.9122, 173.1678)$

(c) Ratio Estimate of Mean

$\mu_y = r\mu_x = \frac{\bar{y}}{\bar{x}}\mu_x = 168.1558$

$\hat{Var}(\hat{\mu_y}) = (1-\frac{n}{N})\frac{s_r^2}{n} = 34.89449$, where $s_r^2 = \frac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1} = 366.3921$

Estimate of μ_y with bound on error of estimation is $\hat{\mu_y} \pm 2\sqrt{\hat{Var}(\hat{\mu_y})} = 168.1558 \pm 2 \times 5.907155 = (156.3415, 179.9701)$

(d) Regression Estimate of Mean

$\mu_y = \bar{y} + b(\mu_x - \bar{x}) = 167.6959$

$\hat{Var}(\hat{\mu_y}) = (1-\frac{n}{N})\frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - a - bx_i)^2}{n-2} = 4.96911$

Estimate of μ_y with bound on error of estimation is $\hat{\mu_y} \pm 2\sqrt{\hat{Var}(\hat{\mu_y})} = 167.6959 \pm 2 \times 2.22915 = (163.2376, 172.1542)$

(e) Difference Estimate of Mean

$\mu_y = \bar{y} + (\mu_x - \bar{x}) = 167.6122$

$\hat{Var}(\hat{\mu_y}) = (1-\frac{n}{N})\frac{1}{n}\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} = 5.42528$, where $\bar{d} = \bar{y} - \bar{x} = 147.89$ and $d_i = y_i - x_i$

Estimate of μ_y with bound on error of estimation is $\hat{\mu_y} \pm 2\sqrt{\hat{Var}(\hat{\mu_y})} = 167.6112 \pm 2 \times 2.329223 = (163.9528, 172.2696)$

(f) Error of Estimation of the Four Estimators

Population mean: $\mu_y = 170.7112$

Error of estimation for SRS estimator: 3.1712

$|\hat{\mu} - \mu_y| = |167.54 - 170.7112| = 3.1712$

Error of estimation for ratio estimator: 2.5554

$|\hat{\mu} - \mu_y| = |168.1558 - 170.7112| = 2.5554$

Error of estimation for regression estimator: 3.0153

$|\hat{\mu} - \mu_y| = |167.6959 - 170.7112| = 3.0153$

Error of estimation for difference estimator: 3.099

$|\hat{\mu} - \mu_y| = |167.6122 - 170.7112| = 3.099$

SRS Estimator > Difference Estimator > Regression Estimator > Ratio Estimator

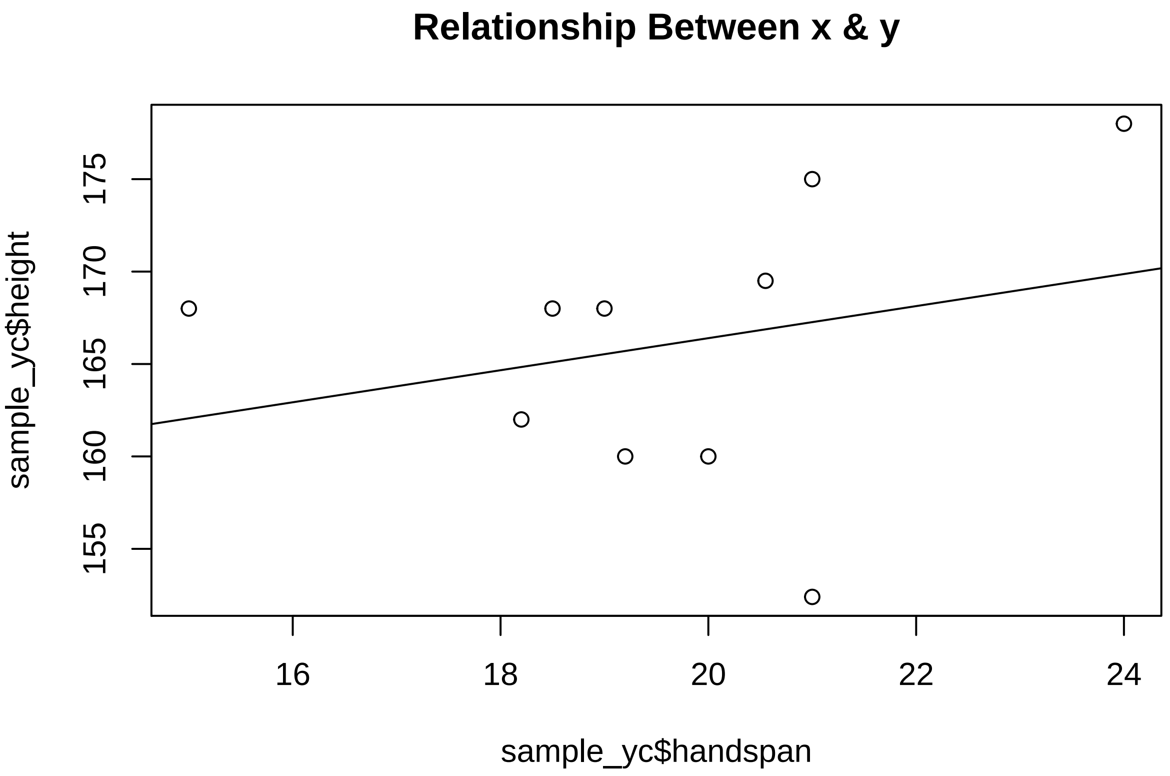
The error of estimation for SRS estimator is the largest, while the error of estimation for ratio estimator is the smallest.

(g) Recommended Estimators

I would recommend using regression estimator. Although the error of estimation for ratio estimator is the smallest, its bound on error of estimation is too wide which makes it less precise. The bound on the error of estimation for regression estimator is the smallest ($B = 4.4582$), which makes it more precise.

Furthermore, the scatterplot below showed that x and y are correlated, but the regression line does not passes through the origin. Thus, using a regression estimator is better.

```r
plot(sample_yc$handspan, sample_yc$height, main= "Relationship Between x & y")
abline(lm(sample_yc$height~sample_yc$handspan))
```



(h) SRS Estimator vs. Other Estimators

No, I do not recommend the SRS estimator over other three estimators. First, its error of estimation is larger than other three estimators. Further, its bound of error of estimation is not smaller than other three estimators (the bound for regression estimator and difference estimator is smaller), which does not make it more precise. Thus, I would not recommend using the SRS estimator.