STA 304H1F-1003H Fall 2019
Assignment 2-Question 3-Solution

Question 3. (18 marks)

Table 1 below gives a summary of some basic statistics for stata 1 (Brand I):

Table 1: Strata 1 Basic statistics

	Stata 1 (Brand I)		
Strata size		N_1 =120	
Strata weight		$W_1 = \frac{N_1}{N}$ =0.4	
Sample size		n_1 =6	
	x		y
Sample Mean	$\bar{x}_1$ =193		$\bar{y}_1$ =202.5
Sample variance	$s^2_{x_1}$ =5132		$s^2_{y_1}$ =6757.5
		$s_{x_1y_1} = \sum_{i=1}^{n_1}(x_i - \bar{x}_1)(y_i - \bar{y}_1)$ =5835	
Correlation coefficient (r)		r_1 =0.991	

Table 2 below gives a summary of some basic statistics for stata 2 (Brand II):

Table 2: Strata 2 Basic statistics

	Strata 2 (Brand II)		
Strata size		N_2 =180	
Strata weight		$W_2 = \frac{N_2}{N}$ =0.6	
Sample size		n_2 =9	
	x		y
Sample Mean	$\bar{x}_2$ =114.111		$\bar{y}_2$ =121.111
Sample Variance	$s^2_{x_2}$ =1788.111		$s^2_{y_2}$ =2242.361
		$s_{x_2y_2} = \sum_{i=1}^{n_2}(x_i - \bar{x}_2)(y_i - \bar{y}_2)$ =1978.611	
Correlation coefficient (r)		r_2 =0.988	

(Part 1) (10 marks) Using Separate Ratio Estimator Method (SR)

(a) (2 marks) Find a basic estimate (without auxiliary information) of the total potential sales. Estimate the variance of your estimator.

The basic estimate is

$$\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2 = (N_1 \times \bar{y}_1) + (N_2 \times \bar{y}_2)$$

From R output:

$$\hat{\tau} = (120 \times 202.5) + (180 \times 121.111) = \boxed{4.61 \times 10^4}$$

(1mk)

The estimated variance is

$$\widehat{\mathbf{V}}(\hat{\tau}) = \widehat{\mathbf{V}}(\hat{\tau}_1) + \widehat{\mathbf{V}}(\hat{\tau}_2) = N_1^2 \times \left(1 - \frac{n_1}{N_1}\right)\frac{s_{y_1}^2}{n_1} + N_2^2 \times \left(1 - \frac{n_2}{N_2}\right)\frac{s_{y_2}^2}{n_2}$$

From R output

$$\widehat{\mathbf{V}}(\hat{\tau}) = (120)^2 \times \left(1 - \frac{6}{120}\right)\frac{6757.5}{6} + (180)^2 \times \left(1 - \frac{9}{180}\right)\frac{2242.361}{9} = \boxed{2.3075975 \times 10^7}$$

(1mk)

(b) **(2 marks) Find a ratio estimate of the total potential sales. Estimate the variance of your estimator.**

The ratio estimate is

$$\hat{\tau}_r^{sr} = N \times \hat{\mu}_{sr} = N\left[W_1 \times \left(\frac{\bar{y}_1}{\bar{x}_1}\right) \times (\mu_{x_1}) + W_2 \times \left(\frac{\bar{y}_2}{\bar{x}_2}\right) \times (\mu_{x_2})\right]$$

From R output:

$$\hat{\tau}_r^{sr} = 300 \times \left[0.4 \times \left(\frac{202.5}{193}\right) \times (204.1666667) + 0.6 \times \left(\frac{121.111}{114.111}\right) \times (117.7777778)\right] = \boxed{4.8206445 \times 10^4}$$

(1mk)

The estimated variance is

$$\widehat{\mathbf{V}}(\hat{\tau}_r^{rs}) = N^2\left[W_1^2 \times \left(1 - \frac{n_1}{N_1}\right)\frac{s_{r_1}^2}{n_1} + W_2^2 \times \left(1 - \frac{n_2}{N_2}\right)\frac{s_{r_2}^2}{n_2}\right]$$

From R output

$$\widehat{\mathbf{V}}(\hat{\tau}_r^{sr}) = (300)^2\left[0.4^2 \times \left(1 - \frac{6}{120}\right)\frac{162.727}{6} + 0.6^2 \times \left(1 - \frac{9}{180}\right)\frac{56.607}{9}\right] = \boxed{5.6461278 \times 10^5}$$

(1mk)

(c) **(2 marks) Find a regression estimate of the total potential sales. Estimate the variance of your estimator.**

The regression coefficients for strata 1 are :

$$b_1 = \frac{s_{x_1y_1}}{s^2_{x_1}} = \frac{5835}{5132} = 1.137 \quad \textbf{and} \quad a_1 = \bar{y}_1 - b_1 \times \bar{x}_1 = 202.5 - 1.137 \times 193 = -16.938$$

The regression coefficients for strata 2 are :

$$b_2 = \frac{s_{x_2y_2}}{s^2_{x_2}} = \frac{1978.611}{1788.111} = 1.107 \quad \textbf{and} \quad a_2 = \bar{y}_2 - b_2 \times \bar{x}_2 = 121.111 - 1.107 \times 114.111 = -5.157$$

The separate regression estimator is then

$$\widehat{\tau}^{sr}_L = N \times \widehat{\mu}^{sr}_L = N[W_1 \times \widehat{\mu}^L_1 + W_2 \times \widehat{\mu}^L_2] = N[W_1 \times (\bar{y}_1 + b_1(\mu_{x_1} - \bar{x}_1)) + W_2 \times (\bar{y}_2 + b_2(\mu_{x_2} - \bar{x}_2))]$$

From R output:

$$\widehat{\tau}^{sr}_L = 300[0.4 \times (202.5 + 1.1369836 \times (204.1666667 - 193)) + 0.6 \times (121.11 + 1.106537 \times (117.7777778 - 114.11))]$$

$$\widehat{\tau}^{sr}_L = \boxed{2.3847612 \times 10^4}$$

(1mk)

The estimated variance is

$$\widehat{\mathbf{V}}(\widehat{\tau}^{rs}_L) = N^2\left[W_1^2 \times \left(1 - \frac{n_1}{N_1}\right)\frac{\mathbf{MSE}_1}{n_1} + W_2^2 \times \left(1 - \frac{n_2}{N_2}\right)\frac{\mathbf{MSE}_2}{n_2}\right]$$

where from R we have that:

$$\mathbf{MSE}_1 = \frac{\sum_{i=1}^{n_1}(y_{1,i} - (a_1 + b_1x_{1,i}))^2}{n_1 - 2} = 333.505, \quad \textbf{and} \quad \mathbf{MSE}_2 = \frac{\sum_{i=1}^{n_2}(y_{2,i} - (a_2 + b_2x_{2,i}))^2}{n_2 - 2} = 106.226$$

and then

$$\widehat{\mathbf{V}}(\widehat{\tau}^{sr}_L) = (300)^2\left[0.4^2 \times (1 - \frac{6}{120})\frac{333.505}{6} + 0.6^2 \times (1 - \frac{9}{180})\frac{106.226}{9}\right] = \boxed{2.5244772 \times 10^5}$$

(1mk)

(d) (3 marks) Compute the relative efficiency of

(i) ratio estimation to basic estimation

The relative efficiency of ratio estimator ($\widehat{\tau}_r^{sr}$) to basic estimator $\widehat{\tau}$ is

$$\widehat{\mathbf{RE}}(\widehat{\tau}_r^{sr}, \widehat{\tau}) = \frac{\widehat{\mathbf{V}}(\widehat{\tau})}{\widehat{\mathbf{V}}(\widehat{\tau}_r^{sr})} = \frac{2.3075975 \times 10^7}{5.64613 \times 10^5} = \boxed{40.87}$$

(1mk)

(ii) ratio estimation to regression estimation

The relative efficiency of ratio estimator ($\widehat{\tau}_r^{sr}$) to regression estimator ($\widehat{\tau}_L^{sr}$) is

$$\widehat{\mathbf{RE}}(\widehat{\tau}_r^{sr}, \widehat{\tau}_L^{sr}) = \frac{\widehat{\mathbf{V}}(\widehat{\tau}_L^{sr})}{\widehat{\mathbf{V}}(\widehat{\tau}_r^{sr})} = \frac{2.5244772 \times 10^5}{5.64613 \times 10^5} = \boxed{0.447}$$

(1mk)

(iii) regression estimation to basic estimation

The relative efficiency of regression estimator ($\widehat{\tau}_L^{sr}$) to basic estimator ($\widehat{\tau}$) is

$$\widehat{\mathbf{RE}}(\widehat{\tau}_L^{sr}, \widehat{\tau}) = \frac{\widehat{\mathbf{V}}(\widehat{\tau})}{\widehat{\mathbf{V}}(\widehat{\tau}_L^{sr})} = \frac{2.3075975 \times 10^7}{2.52448 \times 10^5} = \boxed{91.409}$$

(1mk)

(e) (1 mark) Which method of estimation do you recommend?

From (i) we have:

$$\widehat{\mathbf{RE}}(\widehat{\tau}_r^{sr}, \widehat{\tau}) = \boxed{40.87} > 1 \implies \widehat{\mathbf{V}}(\widehat{\tau}_r^{sr}) < \widehat{\mathbf{V}}(\widehat{\tau})$$

which means that ratio estimator ($\widehat{\tau}_r^{sr}$) is preferable.

From (ii) we have:

$$\widehat{\mathbf{RE}}(\widehat{\tau}_r^{sr}, \widehat{\tau}_L^{sr}) = \boxed{0.447} > 1 \implies \widehat{\mathbf{V}}(\widehat{\tau}_r^{sr}) < \widehat{\mathbf{V}}(\widehat{\tau}_L^{sr})$$

which means that ratio estimator ($\widehat{\tau}_r^{sr}$) is preferable.

From (iii) we have:

$$\widehat{\mathbf{RE}}(\widehat{\tau}_L^{sr}, \widehat{\tau}) = \boxed{91.409} > 1 \implies \widehat{\mathbf{V}}(\widehat{\tau}_L^{sr}) < \widehat{\mathbf{V}}(\widehat{\tau})$$

which means that regression estimator ($\widehat{\tau}_L^{sr}$) is preferable.

Based on the above three conclusions, we would recommand separate ratio estimator

(1mk)

(Part 2) (6 marks) Using Combined Ratio Estimator Method (CR)

(a) (2 marks) Find a basic estimate (without auxiliary information) of the mean potential sales. Estimate the variance of your estimator.

The estimate of μ_y is

$$\overline{y}_{st} = W_1 \times \bar{y}_1 + W_2 \times \bar{y}_2 = 0.4 \times 202.5 + 0.6 \times 121.111 = \boxed{153.667}$$

(1mk)

The estimated variance is

$$\widehat{\mathbf{V}}(\overline{y}_{st}) = W_1^2 \times \left(1 - \frac{n_1}{N_1}\right)\frac{s_{y1}^2}{n_1} + W_2^2 \times \left(1 - \frac{n_2}{N_2}\right)\frac{s_{y2}^2}{n_2}$$

$$\widehat{\mathbf{V}}(\overline{y}_{st}) = 0.4^2 \times (1 - \frac{6}{120})\frac{6757.5}{6} + 0.6^2 \times (1 - \frac{9}{180})\frac{2242.361}{9} = \boxed{256.4}$$

(1mk)

(b) (2 marks) Find a ratio estimate of the mean potential sales. Estimate the variance of your estimator.

The true value of μ_x is

$$\mu_x = W_1 \times \mu_{x_1} + W_2 \times \mu_{x_2} = 0.4 \times 204.167 + 0.6 \times 117.778 = 152.333$$

The estimate of μ_y using the stratified random sample is

$$\overline{y}_{st} = W_1 \times \bar{y}_1 + W_2 \times \bar{y}_2 = 0.4 \times 202.5 + 0.6 \times 121.111 = 153.667$$

The estimate of μ_x using the stratified random sample is

$$\overline{x}_{st} = W_1 \times \bar{x}_1 + W_2 \times \bar{x}_2 = 0.4 \times 193 + 0.6 \times 114.111 = 145.667$$

The combined ratio estimate of μ is

$$\widehat{\mu}_{cr} = \frac{\overline{y}_{st}}{\overline{x}_{st}} \times (\mu_x) = \frac{153.667}{145.667} \times (152.3333333) = \boxed{160.699}$$

(1mk)

The estimated variance is

$$\widehat{\mathbf{V}}(\widehat{\mu}_{cr}) = W_1^2 \times \left(1 - \frac{n_1}{N_1}\right)\frac{s_{cr,1}^2}{n_1} + W_2^2 \times \left(1 - \frac{n_2}{N_2}\right)\frac{s_{cr,2}^2}{n_2}$$

where

$$s_{cr,1}^2 = \frac{\sum_{j=1}^{n_1}(y_j - \hat{r}_{cr}x_j)^2}{n_1 - 1} = 159.213$$

and

$$s_{cr,2}^2 = \frac{\sum_{j=1}^{n_2}(y_j - \hat{r}_{cr}x_j)^2}{n_2 - 1} = 58.323$$

and then

$$\widehat{\mathbf{V}}(\widehat{\mu}_{cr}) = 0.4^2 \times (1 - \frac{6}{120})\frac{159.213}{6} + 0.6^2 \times (1 - \frac{9}{180})\frac{58.323}{9} = \boxed{6.25}$$

(1mk)

(c) (1 mark) Compute the relative efficiency of ratio estimation to basic estimation

The relative efficiency of ratio estimator $(\widehat{\mu}_r^{sr})$ to basic estimator $\widehat{\mu}$ is

$$\widehat{\mathrm{RE}}(\widehat{\mu}_r^{cr}, \widehat{\mu}) = \frac{\widehat{\mathbf{V}}(\widehat{\mu})}{\widehat{\mathbf{V}}(\widehat{\mu}_r^{cr})} = \frac{256.4}{6} = \boxed{41.026}$$

(1mk)

(d) (1 mark) Which method of estimation do you recommend?

$$\widehat{\mathrm{RE}}(\widehat{\mu}_r^{cr}, \widehat{\mu}) = \frac{\widehat{\mathbf{V}}(\widehat{\mu})}{\widehat{\mathbf{V}}(\widehat{\mu}_r^{cr})} = \frac{256.4}{6} = \boxed{41.026} > 1 \implies \widehat{\mathbf{V}}(\widehat{\mu}_r^{sr}) < \widehat{\mathbf{V}}(\widehat{\mu})$$

which means that combined ratio estimator $(\widehat{\mu}_r^{sr})$ is preferable.

(1mk)

R code

```
# strata 1 (brand I)
N1<-120
tot_x1<-24500
mu_x1<-tot_x1/N1
x1<-c(204,143,82,256,275,198)
y1<-c(210,160,75,280,300,190)
n1<-length(y1)

# strata 2 (brand II)
N2<-180
tot_x2<-21200
mu_x2<-tot_x2/N1
x2<-c(137,189,119,63,103,107,159,63,87)
y2<-c(150,200,125,60,110,100,180,75,90)
n2<-length(y2)

#  basic statistics

# Population size
N<-N1+N2

# Strata weights
W1<-N1/N
W2<-N2/N

# sampling fractions for each strata
f1<-n1/N1
f2<-n2/N2

# strata samples means for x and y
y1bar<-mean(y1)
y2bar<-mean(y2)

x1bar<-mean(x1)
x2bar<-mean(x2)

# strata samples variances for x and y
s2y1<-var(y1)
s2y2<-var(y2)

s2x1<-var(x1)
s2x2<-var(x2)

# strata samples cross products between x and y
sx1y1<-sum( (x1-mean(x1))*(y1-mean(y1)) )/(n1-1)
sx2y2<-sum( (x2-mean(x2))*(y2-mean(y2)) )/(n2-1)

# strata samples corelation coefficients between x and y
cor1<-cor(y1,x1)
cor2<-cor(y2,x2)

# some verifications:
```

```r
# r1 and sx1y1/(sqrt( s2x1*s2y1)) should be the same
# r2 and sx2y2/(sqrt( s2x2*s2y2)) should be the same


# (Part 1)  Separate ratio estimation (sr)

# (a) Basic method
mu1hat<-mean(y1)
var_mu1hat<-(1-f1)*(s2y1/n1)

mu2hat<-mean(y2)
var_mu2hat<-(1-f2)*(s2y2/n2)

mu_b_hat_sr<-W1*mu1hat + W2*mu2hat
var_mu_b_hat_sr<-W1^2*var_mu1hat+W2^2*var_mu2hat

tau_b_hat_sr<-N*mu_b_hat_sr
var_tau_b_hat_sr<-N^2*var_mu_b_hat_sr


# b) Ratio
r1<-y1bar/x1bar
r2<-y2bar/x2bar

s2r1<-var(y1-r1*x1)
s2r2<-var(y2-r2*x2)

mu1_r_hat<-r1*mu_x1
mu2_r_hat<-r2*mu_x2

var_mu1_r_hat<-(1-f1)*(s2r1/n1)
var_mu2_r_hat<-(1-f2)*(s2r2/n2)

tau1_r_hat<-N1*mu1_r_hat
tau2_r_hat<-N2*mu2_r_hat

var_tau1_r_hat<-N1^2*var_mu1_r_hat
var_tau2_r_hat<-N2^2*var_mu2_r_hat

mu_r_hat_sr<-W1*mu1_r_hat + W2*mu2_r_hat
var_mu_r_hat_sr<-W1^2*var_mu1_r_hat + W2^2*var_mu2_r_hat

tau_r_hat_sr<-N*mu_r_hat_sr
var_tau_r_hat_sr<-N^2*var_mu_r_hat_sr

# c) regression
b1<-sx1y1/s2x1
b2<-sx2y2/s2x2

a1<-y1bar - b1*x1bar
a2<-y2bar - b2*x2bar

y1hat<-a1+b1*y1
```

```r
y2hat<-a2+b2*y2

e1<-y1-y1hat
e2<-y2-y2hat

MSE1<-sum(e1^2)/(n1-2)
MSE2<-sum(e2^2)/(n2-2)

mu1_L_hat<-y1bar + b1*(mu_x1-x1bar)
mu2_L_hat<-y2bar + b2*(mu_x2-x2bar)

var_mu1_L_hat<-(1-f1)*(MSE1/n1)
var_mu2_L_hat<-(1-f2)*(MSE2/n2)

tau1_L_hat<-N1*mu1_L_hat
tau2_L_hat<-N2*mu2_L_hat

var_tau1_L_hat<-N1^2*var_mu1_L_hat
var_tau2_L_hat<-N2^2*var_mu2_L_hat

mu_L_hat_sr<-W1*mu1_L_hat + W2*mu2_L_hat
var_mu_L_hat_sr<-W1^2*var_mu1_L_hat + W2^2*var_mu2_L_hat

tau_L_hat_sr<-W1*tau1_L_hat + W2*tau2_L_hat
var_tau_L_hat_sr<-W1^2*var_tau1_L_hat + W2^2*var_tau2_L_hat


#  d) relative efficiency

# (i)
eff_r_to_b<-var_tau_b_hat_sr/var_tau_r_hat_sr

# (ii)
eff_r_to_L<-var_tau_L_hat_sr/var_tau_r_hat_sr

# (iii)
eff_L_to_b<-var_tau_b_hat_sr/var_tau_L_hat_sr

# (Part 2) composed ratio method

# (a)
mu_b_hat_cr<-W1*y1bar + W2*y2bar
var_mu_b_hat_cr<-W1^2*(1-f1)*(s2y1/n1) + W2^2*(1-f2)*(s2y2/n2)

# (b)
mu_x<-W1*mu_x1 + W2*mu_x2
muy_str_hat<-W1*y1bar + W2*y2bar
mux_str_hat<-W1*x1bar + W2*x2bar

r_cr<-muy_str_hat/mux_str_hat

mu_r_hat_cr<-r_cr*(mu_x)
```

```r
s2cr1<-sum( (y1-r_cr*x1)^2 )/(n1-1)
s2cr2<-sum( (y2-r_cr*x2)^2 )/(n2-1)

var_mu_r_hat_cr<-W1^2*(1-f1)*(s2cr1/n1) + W2^2*(1-f2)*(s2cr2/n2)

# (c) relative efficiency RE
RE_cr<-var_mu_b_hat_cr/var_mu_r_hat_cr
```