Semester 2 Applied Statistics 2015

Solution to Tutorial 11

1. We have $p_1 = \dfrac{6}{24} = 0.25$, $p_2 = \dfrac{3}{60} = 0.05$ and $p_3 = \dfrac{5}{40} = 0.125$.

   (a) They are $N_1 = 25,100 \times 0.15 = 3,765$, $N_2 = 25,100 \times 0.5 = 12,550$ and $25,100 \times 0.35 = 8,785$ respectively for the three age groups.

   (b) The estimate of the overall unemployment rate in the district and the standard error estimate are

$$
\begin{aligned}
\widehat{P}_{st} &= \frac{N_1}{N}p_1 + \frac{N_2}{N}p_2 + \frac{N_2}{N}p_3 \\
&= (0.15)(.25) + (0.5)(.05) + (0.35)(.125) = 0.10625 \\
\text{var}(\widehat{P}_{st}) &= \left(\frac{N_1}{N}\right)^2 \left(1 - \frac{n_1}{N_1}\right) \frac{p_1(1-p_1)}{n_1 - 1} + \left(\frac{N_2}{N}\right)^2 \left(1 - \frac{n_2}{N_2}\right) \frac{p_2(1-p_2)}{n_2 - 1} \\
&\quad + \left(\frac{N_3}{N}\right)^2 \left(1 - \frac{n_3}{N_3}\right) \frac{p_3(1-p_3)}{n_3 - 1} \\
&= 0.15^2 \left(1 - \frac{24}{3,765}\right) \frac{.25(.75)}{23} + 0.5^2 \left(1 - \frac{60}{12,550}\right) \frac{.05(.95)}{59} \\
&\quad + 0.35^2 \left(1 - \frac{40}{8,785}\right) \frac{.125(.875)}{39} \\
&= 0.0007245 \\
\text{se}(\widehat{P}_{st}) &= \sqrt{0.0007245} = 0.0269
\end{aligned}
$$

   (c) The estimate of the overall unemployment rate in the district using post-stratification is still $\widehat{P}_{pst} = \widehat{P}_{st} = 0.10625$ and the standard error estimate is

$$
\begin{aligned}
\text{var}(\widehat{P}_{pst}) &= \frac{1}{n}\left(1 - \frac{n}{N}\right) \sum_{l=1}^{L} W_l p_l (1 - p_l) \\
&= \frac{1}{124}\left(1 - \frac{124}{25,100}\right) [.15 \times .25(.75) + .5 \times .05(.95) + .35 \times .125(.875)] \\
&= 0.00074 \\
\text{se}(\widehat{P}_{pst}) &= \sqrt{0.00074} = 0.0272
\end{aligned}
$$

   (d) Since the sampling proportions $\dfrac{n_1}{n} = 0.194$, $\dfrac{n_2}{n} = 0.484$ and $\dfrac{n_3}{n} = 0.323$ are quite close to the population proportions $\dfrac{N_1}{N} = 0.15$, $\dfrac{N_2}{N} = 0.5$ and $\dfrac{N_3}{N} = 0.35$, the use of post-stratification to adjust the sample proportions to population proportions is not necessary.

(e) The total sample size n is

$$\begin{aligned}
\sum_{l=1}^{L} W_l\sqrt{p_l(1-p_l)} &= W_1\sqrt{p_1(1-p_1)} + W_2\sqrt{p_2(1-p_2)} + W_3\sqrt{p_3(1-p_3)} \\
&= 0.15\sqrt{.25(.75)} + 0.5\sqrt{.05(.95)} + 0.35\sqrt{.125(.875)} \\
&= 0.064952 + 0.108972 + 0.115752 = 0.289676 \\
\sum_{l=1}^{L} W_l p_l(1-p_l) &= W_1p_1(1-p_1) + W_2p_2(1-p_2) + W_3p_3(1-p_3) \\
&= 0.15(.25)(.75) + 0.5(.05)(.95) + 0.35(.125)(.875) \\
&= 0.028125 + 0.02375 + 0.038281 = 0.091056 \\
V &= \left(\frac{\delta_\mu}{1.96}\right)^2 \approx \left(\frac{.02}{2}\right)^2 = .0001 \\
n &= \frac{\left(\sum_{l=1}^{L} W_l\sqrt{p_l(1-p_l)}\right)^2}{V + \frac{1}{N}\sum_{l=1}^{L} W_l p_l(1-p_l)} = \frac{0.289676^2}{0.0001 + \frac{1}{25100}0.091056} \\
&= 809.7464 \text{ or } 810
\end{aligned}$$

$$\begin{aligned}
n_1 &= n\left(\frac{W_1\sqrt{p_1(1-p_1)}}{\sum_{i=1}^{L} W_i\sqrt{p_i(1-p_i)}}\right) = 810 \times \frac{0.064952}{0.289676} = 181.62 = 181 \\
n_2 &= n\left(\frac{W_2\sqrt{p_2(1-p_2)}}{\sum_{i=1}^{L} W_i\sqrt{p_i(1-p_i)}}\right) = 810 \times \frac{0.108972}{0.289676} = 304.71 = 305 \\
n_3 &= n\left(\frac{W_3\sqrt{p_3(1-p_3)}}{\sum_{i=1}^{L} W_i\sqrt{p_i(1-p_i)}}\right) = 810 \times \frac{0.115752}{0.289676} = 323.67 = 324
\end{aligned}$$

2. $\bar{X} = \dfrac{X}{N} = \dfrac{20500}{256} = 80.07813$

(a) 1. Ratio estimate and its s.e. for the percentage increase in monthly sales after the launch of the promotion campaign:

$$\begin{aligned}
\widehat{R} &= \frac{\sum_i y_i}{\sum_i x_i} = \frac{821}{798} = 1.028822 \\
s_r^2 &= \frac{1}{n-1}\left(\sum_i y_i^2 - 2\widehat{R}\sum_i x_i y_i + \widehat{R}^2\sum_i x_i^2\right) \\
&= \frac{1}{9}\left[75,765 - 2 \times \frac{821}{798} \times 71,672 + \left(\frac{821}{798}\right)^2 \times 71,028\right] = 385.6534 \\
\mathrm{var}(\widehat{R}) &= \frac{1}{\bar{X}^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} = \frac{1}{80.07813^2}\left(1 - \frac{10}{256}\right)\frac{385.6534}{10} = 0.005779 \\
\mathrm{se}(\widehat{R}) &= \sqrt{0.005779} = 0.076021
\end{aligned}$$

Hence the percentage increase is 2.88%.

2. For the estimator $\widehat{R}_o = \frac{\bar{y}}{\bar{X}}$, the estimate and its s.e. for the percentage increase

in monthly sales after the launch of the promotion campaign:

$$\begin{aligned}
\widehat{R}_o &= \frac{\bar{y}}{\overline{X}} = \frac{821/10}{80.07813} = 1.025249 \\
s_y^2 &= \frac{1}{n-1}\left[\sum_i y_i^2 - \left(\sum_i y_i\right)^2 /n\right] \\
&= \frac{1}{9}\left[75,765 - 821^2/10\right] = 928.9889 \\
\mathrm{var}(\widehat{R}_o) &= \frac{1}{\overline{X}^2}\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} = \frac{1}{80.07813^2}\left(1-\frac{10}{256}\right)\frac{928.9889}{10} = 0.013921 \\
\mathrm{se}(\widehat{R}_o) &= \sqrt{0.013921} = 0.117988
\end{aligned}$$

Hence the percentage increase is 2.52%.

(b) The two estimates and their s.e. for the average monthly sales after the launch of the promotion campaign:

$$\begin{aligned}
\widehat{\overline{Y}}_r &= \overline{X}\widehat{R} = 80.07813 \times 1.028822 = 82.38614 \\
\mathrm{se}(\widehat{\overline{Y}}_r) &= \overline{X} \times \mathrm{se}(\widehat{R}) = 80.07813 \times 0.076021 = 6.087601 \\
\widehat{\overline{Y}} &= \overline{y} = 82.1 \\
\mathrm{se}(\widehat{\overline{Y}}) &= \overline{X} \times \mathrm{se}(\widehat{R}') = 80.07813 \times 0.117988 = 9.448281
\end{aligned}$$

(c) New estimate for the average monthly sales after the launch of the promotion campaign:

$$\begin{aligned}
\widehat{\overline{Y}}_d &= \overline{X} + \bar{d} = 80.07813 + 2.3 = 82.37813 \\
\mathrm{var}(\widehat{\overline{Y}}_d) &= \left(1-\frac{n}{N}\right)\frac{s_d^2}{n} = \left(1-\frac{10}{256}\right)\frac{377.3444}{10} = 36.26043844 \\
\mathrm{se}(\widehat{\overline{Y}}_d) &= \sqrt{36.26043844} = 6.021664092
\end{aligned}$$

(d) Estimators $\widehat{\overline{Y}}$ and $\widehat{\overline{Y}}_d$ are unbiased but estimator $\widehat{\overline{Y}}_r$ is only approximately unbiased.

$$E(\bar{X} + \bar{d}) = \bar{X} + E(\bar{d}) = \bar{X} + \bar{Y} - \bar{X} = \bar{Y} \quad \text{unbiased}$$

Estimator $\widehat{\overline{Y}}_r$ assumes a zero intercept for the relationship between x and y. Estimator $\widehat{\overline{Y}}_d$ assumes a unit slope for the relationship between x and y. It is known that $\mathrm{var}(\widehat{\overline{Y}}) > \mathrm{var}(\widehat{\overline{Y}}_r)$, if

$$\hat{\rho} > \frac{\bar{y}}{\bar{x}}\frac{s_x}{2s_y} = \frac{82.1}{79.8}\frac{28.57271}{2(30.47932)} = 0.482233.$$

Also $\mathrm{var}(\widehat{\overline{Y}}) > \mathrm{var}(\widehat{\overline{Y}}_d)$, if

$$s_y^2 > s_d^2 \;\Rightarrow\; s_y^2 > s_y^2 - 2\hat{\rho}s_x s_y + s_x^2 \Rightarrow \hat{\rho} > \frac{s_x}{2s_y} = \frac{28.57271}{2(30.47932)} = 0.468723.$$

These conditions are satisfied since $\hat{\rho} = 0.7854$. Results show that $\mathrm{se}(\widehat{\overline{Y}}_r) \simeq \mathrm{se}(\widehat{\overline{Y}}_d) < \mathrm{se}(\widehat{\overline{Y}})$.

3. (a) We have

$$\bar{y} = \frac{\sum_i y_i}{n} = \frac{177}{10} = 17.7,\ \bar{x} = \frac{\sum_i x_i}{n} = \frac{179}{10} = 17.9 \text{ and } \overline{X} = \frac{X}{N} = \frac{16,300}{1,000} = 16.3.$$

The ordinary estimate and ratio estimate of the average number of man-hours lost Y because of sickness this year are respectively:

$$\widehat{\overline{Y}} = \bar{y} = 17.7 \quad \text{and} \quad \widehat{\overline{Y}}_r = \overline{X}\frac{\bar{y}}{\bar{x}} = 16.3 \times \frac{17.7}{17.9} = 16.12$$

The discrepancy lies in the ratio of the population mean to sample mean,

$$\frac{\overline{X}}{\bar{x}} = \frac{16.3}{17.9} = 0.911$$

which is quite close to 1 showing that the discrepancy is not large.

(b) Ratio estimator for the total number of man-hours lost Y because of sickness this year:

$$\begin{aligned}
\widehat{Y}_r &= Xr = X \times \frac{\sum_i y_i}{\sum_i x_i} = 16300 \times \frac{177}{179} = 16300(0.9888) = 16117.877 \\
s_r^2 &= \frac{1}{n-1}\left(\sum_i y_i^2 - 2r\sum_i x_i y_i + r^2 \sum_i x_i^2\right) \\
&= \frac{1}{9}\left[3,843 - 2 \times \frac{177}{179} \times 3,927 + \left(\frac{177}{179}\right)^2 \times 4,067\right] = 5.9310 \\
\text{se}(\widehat{Y}_r) &= N\sqrt{\left(1-\frac{n}{N}\right)\frac{s_r^2}{n}} = 1,000\sqrt{\left(1-\frac{10}{1,000}\right)\frac{5.9310}{10}} = \sqrt{587,171.562} \\
&= 766.2712
\end{aligned}$$

(c) Regression estimator for the total number of man-hours lost Y because of sickness this year:

$$\begin{aligned}
\sum_i x_i y_i - n\overline{xy} &= 3,927 - 10 \times 17.9 \times 17.7 = 758.7 \\
\sum_i x_i^2 - n\overline{x}^2 &= 4,067 - 10 \times 17.9^2 = 862.9, \\
\sum_i y_i^2 - n\overline{y}^2 &= 3,843 - 10 \times 17.7^2 = 710.1.
\end{aligned}$$

We have

$$\begin{aligned}
s_y^2 &= \frac{1}{n-1}(\sum_i y_i^2 - n\overline{y}^2) = \frac{710.1}{9} = 78.9 \\
\hat{\rho} &= \frac{\sum_i x_i y_i - n\overline{xy}}{\sqrt{(\sum_i x_i^2 - n\overline{x}^2)(\sum_i y_i^2 - n\overline{y}^2)}} = \frac{758.7}{\sqrt{862.9 \times 710.1}} = 0.9692 \\
s_{reg}^2 &= s_y^2(1-\hat{\rho}^2) = 78.9(1-0.9692^2) = 4.7854, \\
b &= \frac{\sum_i x_i y_i - n\overline{xy}}{\sum_i x_i^2 - n\overline{x}^2} = \frac{758.7}{862.9} = 0.8792 \\
\widehat{Y}_{reg} &= N[\overline{y} + b(\overline{X} - \overline{x})] \\
&= 1,000\left[17.7 + 0.8792\left(\frac{16,300}{1,000} - 17.9\right)\right] = 16,293.209 \\
\text{se}(\widehat{Y}_{reg}) &= 1,000\sqrt{\left(1-\frac{n}{N}\right)\frac{s_{reg}^2}{n}} \\
&= 1,000\sqrt{\left(1-\frac{10}{1,000}\right)\frac{4.7854}{10}} = \sqrt{473189.9408} = 687.888 < 766.2712.
\end{aligned}$$

The estimator $\widehat{Y}_{reg}$ is more efficient than $\widehat{Y}_r$ because the assumption that the linear relationship between x and y passes through the origin can be dropped. Since $\hat{\rho} = 0.9692$ is positive and close to 1 which shows a strong and positive relationship between x and y, the use of x as an auxiliary variable is suitable.

(d) The rate of change of man-hours lost is

$$\text{Rate of change} = \frac{\text{Lost in current year} - \text{lost in last year}}{\text{Lost in last year}} = \frac{\text{Lost in current year}}{\text{Lost in last year}} - 1.$$

Hence the ratio estimator is $R' = R - 1$ and hence s_r^2 in part (a) is used for S^2. The error bound for R' is $\delta_{r'} = 0.05$. Hence the error bound for R is $\delta_r = 0.05$ and the error bound for the mean estimate is

$$\delta_{\mu_r} = \overline{X}(\delta_r) = \frac{16,300}{1,000} \times 0.05 = 0.815,$$

i.e. $\delta_\mu = 0.815$.

$$n = \frac{NS_r^2}{N(\delta_\mu)^2/z_{\alpha/2}^2 + S_r^2} = \frac{1,000 \times 5.931}{1,000 \times \dfrac{0.815^2}{1.96^2} + 5.931} = 33.16.$$

Take $n = 34$.

4. To show that

$$\text{var}(\widehat{\overline{Y}}_{st,ds}) = \sum_{l=1}^{L} \frac{1}{n_l}\widehat{W}_l^2 s_l^2 + \sum_{l=1}^{L}\frac{1}{n'}\widehat{W}_l(\bar{y}_l - \bar{y}_{st,ds})^2$$

where $\widehat{W}_l = w_l = \dfrac{n_l'}{n'}$ and $\widehat{\overline{Y}}_{st,ds} = \displaystyle\sum_{l=1}^{L} w_l \bar{y}_l$, we note that the first sample has size n' and the second samples are random subsamples of size $n_l = n_l' f_l$ where the sampling fractions f_l in the second phase are fixed.

Firstly, we show that

$$\mathrm{Var}(\widehat{\widehat{Y}}_{st,ds}) = \mathrm{Var}\left[E\left(\sum_{l=1}^{L} w_l\bar{y}_l|w_l\right)\right]+E\left[\mathrm{Var}\left(\sum_{l=1}^{L} w_l\bar{y}_l|w_l\right)\right] = \left(1-\frac{n'}{N}\right)\frac{S^2}{n'}+\sum_l \frac{W_lS_l^2}{n'}\left(\frac{1}{f_l}-1\right)$$

where S^2 is the population variance. Suppose y_{li} were measured on all n_l' first sample in stratum l, not just the subsample of size n_l. Note that

$$E\left(\sum_{l=1}^{L} w_l\bar{y}_l|w_l\right) = \sum_{l=1}^{L} w_l\bar{y}_l' = \bar{y}'$$

the mean of a SRS of size n' from the population and hence

$$\mathrm{Var}\left[E\left(\sum_{l=1}^{L} w_l\bar{y}_l|w_l\right)\right] = \mathrm{Var}\left(\sum_{l=1}^{L} w_l\bar{y}_l'\right) = \mathrm{Var}(\bar{y}') = \left(1-\frac{n'}{N}\right)\frac{S^2}{n'}.$$

Moreover

$$\begin{aligned}\mathrm{Var}\left(\sum_{l=1}^{L} w_l\bar{y}_l|w_l\right) &= \sum_{l=1}^{L} w_l^2\left(1-\frac{n_l}{n_l'}\right)\frac{S_l^2}{n_l} = \sum_{l=1}^{L} w_l^2S_l^2\left(\frac{1}{n_l}-\frac{1}{n_l'}\right) = \sum_{l=1}^{L} w_l^2S_l^2\left(\frac{1}{w_ln'f_l}-\frac{1}{w_ln'}\right)\\ &= \sum_{l=1}^{L}\frac{w_lS_l^2}{n'}\left(\frac{1}{f_l}-1\right)\end{aligned}$$

since $n_l' = w_ln'$ and $n_l = n_l'f_l = w_ln'f_l$. Hence

$$E\left[\mathrm{Var}\left(\sum_{l=1}^{L} w_l\bar{y}_l|w_l\right)\right] = E\left[\sum_{l=1}^{L}\frac{w_lS_l^2}{n'}\left(\frac{1}{f_l}-1\right)\right] = \sum_{l=1}^{L}\frac{W_lS_l^2}{n'}\left(\frac{1}{f_l}-1\right)$$

Hence

$$\mathrm{Var}(\widehat{\widehat{Y}}_{st,ds}) = \left(1-\frac{n'}{N}\right)\frac{S^2}{n'}+\sum_{l=1}^{L}\frac{W_lS_l^2}{n'}\left(\frac{1}{f_l}-1\right).$$

Secondly, we have

$$(N-1)S^2 = \sum_{l=1}^{L}(N_l-1)S_l^2+\sum_{l=1}^{L}N_l(\bar{Y}_l-\bar{Y})^2$$

by the analysis of variance. Multiplying both sides by $\dfrac{N-n'}{n'N(N-1)}$, we have

$$\frac{(N-n')S^2}{n'N} = S^2\left(\frac{1}{n'}-\frac{1}{N}\right) = \frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}\left(W_l-\frac{1}{N}\right)S_l^2+\frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}W_l(\bar{Y}_l-\bar{Y})^2$$

Hence

$$\begin{aligned}
& \mathrm{Var}(\widehat{\widehat{\overline{Y}}}_{st,ds}) \\
= & \frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}\left(W_l-\frac{1}{N}\right)S_l^2+\frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}W_l(\bar{Y}_l-\bar{Y})^2+\sum_{l=1}^{L}\frac{W_lS_l^2}{n'}\left(\frac{1}{f_l}-1\right) \\
= & \sum_{l=1}^{L}W_lS_l^2\left(\frac{1}{n'f_l}-\frac{1}{n'}\right)+\frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}\left(W_lS_l^2-\frac{S_l^2}{N}\right)+\frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}W_l(\bar{Y}_l-\bar{Y})^2 \\
= & \sum_{l=1}^{L}W_lS_l^2\left(-\frac{1}{n'}+\frac{N-n'}{n'(N-1)}+\frac{1}{n'f_l}-\frac{N-n'}{n'(N-1)W_lN}\right)+\frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}W_l(\bar{Y}_l-\bar{Y})^2 \\
= & \sum_{l=1}^{L}W_lS_l^2\left(-\frac{1}{N}+\frac{N-n'}{n'N(N-1)}+\frac{1}{n'f_l}-\frac{N-n'}{n'(N-1)W_lN}\right)+\frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}W_l(\bar{Y}_l-\bar{Y})^2 \\
= & \sum_{l=1}^{L}W_lS_l^2\left(\frac{1}{n'f_l}-\frac{1}{N}\right)+\frac{N-n'}{n'N(N-1)}\sum_{l=1}^{L}(W_l-1)\,S_l^2+\frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}W_l(\bar{Y}_l-\bar{Y})^2 \\
\simeq & \sum_{l=1}^{L}W_lS_l^2\left(\frac{1}{n'f_l}-\frac{1}{N}\right)+\frac{N-n'}{n'(N-1)}\sum_{l=1}^{L}W_l(\bar{Y}_l-\bar{Y})^2 \quad (\text{when } \tfrac{N-n'}{n'N(N-1)}\approx 0) \\
\simeq & \sum_{l=1}^{L}\frac{W_l^2S_l^2}{n_l}+\frac{1}{n'}\sum_{l=1}^{L}W_l(\bar{Y}_l-\bar{Y})^2 \quad (\text{when } \tfrac{N-n'}{N-1}\approx 1 \text{ and } W_ln'f_l=n_l'f_l=n_l)
\end{aligned}$$

Note

$$\begin{aligned}
-\frac{1}{n'}+\frac{N-n'}{n'(N-1)} &= \frac{-N+1+N-n'}{n'(N-1)}=\frac{N-n'N}{n'N(N-1)}=\frac{N-n'N+n'-n'}{n'N(N-1)} \\
&= \frac{-n'(N-1)+(N-n')}{n'N(N-1)}=\frac{-n'+\frac{(N-n')}{N-1}}{n'N}=-\frac{1}{N}+\frac{N-n'}{n'N(N-1)}.
\end{aligned}$$

Hence

$$\mathrm{var}(\widehat{\widehat{\overline{Y}}}_{st,ds})\simeq\sum_{l=1}^{L}\frac{\widehat{W}_l^2S_l^2}{n_l}+\frac{1}{n'}\sum_{l=1}^{L}\widehat{W}_l(\bar{Y}_l-\bar{Y})^2$$

Extra exercise

1. Summary statistics:

$$\sum_i y_i=34.1;\ \sum_i x_i=1707;\ \sum_i y_i^2=117.67;\ \sum_i x_iy_i=5,813.4;\sum_i x_i^2=292,069.$$

   (a) Standard deviation of X and Y:

$$\begin{aligned}
s_x &= \sqrt{\frac{1}{n-1}(\sum_i x_i^2-n\overline{x}^2)}=\sqrt{\frac{1}{9}(292,069-10\times 170.7^2)}=8.718435 \\
s_y &= \sqrt{\frac{1}{n-1}(\sum_i y_i^2-n\overline{y}^2)}=\sqrt{\frac{1}{9}(117.67-10\times 34.1^2)}=0.392853
\end{aligned}$$

(b) Condition for efficient estimates:

$$\begin{aligned}
\sum_{i\in\mathcal{S}} x_i y_i - n\overline{xy} &= 5,813.4 - 10 \times 170.7 \times 3.41 = -0.83 \\
\sum_{i\in\mathcal{S}} x_i^2 - n\overline{x}^2 &= 292,069 - 10 \times 170.7^2 = 76.0\dot{1}, \\
\sum_{i\in\mathcal{S}} y_i^2 - n\overline{y}^2 &= 117.67 - 10 \times 3.41^2 = 0.154\dot{3}.
\end{aligned}$$

We have

$$\begin{aligned}
\hat{\rho} &= \frac{\sum_{i\in\mathcal{S}} x_i y_i - n\overline{xy}}{\sqrt{(\sum_{i\in\mathcal{S}} x_i^2 - n\overline{x}^2)(\sum_{i\in\mathcal{S}} y_i^2 - n\overline{y}^2)}} = \frac{-0.83}{\sqrt{76.0\dot{1} \times 0.154\dot{3}}} = -0.242331 \\
\frac{cv(x)}{2cv(y)} &= \frac{\overline{y}s_x}{2\overline{x}s_y} = \frac{3.41 \times 8.718435}{2 \times 170.7 \times 0.392853} = 0.221666121 > \hat{\rho} = -0.242331
\end{aligned}$$

Since $\hat{\rho} < \dfrac{cv(x)}{2cv(y)}$ and $\hat{\rho} = -0.242331$ is negative and close to zero, the ratio estimate is not preferred.

(c) To estimate the mean score of some measure of the function of lung, we should use ordinary estimator of mean $\overline{Y}$:

$$\begin{aligned}
\widehat{\overline{Y}} &= \overline{y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{34.1}{10} = 3.41 \\
\text{se}(\widehat{\overline{Y}}) &= \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}} = \sqrt{\left(1 - \frac{10}{560}\right)\frac{0.392853^2}{10}} = 0.1231
\end{aligned}$$

2. Simple random sample with $N = 832$ and $n = 8$.

(a) We have $\bar{y} = 24.483$. The total amount spent on home loan payment using ordinary estimator and the standard error estimate:

$$\begin{aligned}
\widehat{Y} &= N\bar{y} = 832 \times 24.483 = 20370.06 \\
s_y^2 &= \frac{1}{n-1}\left(\sum_i y_i^2 - n\bar{y}^2\right) = \frac{1}{7}(5439.616 - 8 \times 24.483^2) = 92.0257 \\
\text{var}(\widehat{Y}) &= N^2\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = 832^2\left(1 - \frac{8}{832}\right)\frac{92.0257}{8} = 7,886,232 \\
\text{se}(\widehat{Y}) &= \sqrt{7,886,232} = 2808.244
\end{aligned}$$

(b) When X_1 the household income is used, we have $\bar{X}_1 = 54.5$ and $\bar{x}_1 = 24.483$. The ratio estimators of the total amount of home loan payment is:

$$\begin{aligned}
\widehat{R}_1 &= r_1 = \frac{\sum_i y_i}{\sum_i x_{1i}} = \frac{195.866}{447.7} = 0.4375 \\
\widehat{Y}_{r1} &= \widehat{R}_1 N\overline{X}_1 = 0.4375 \times 832 \times 54.5 = 19837.72
\end{aligned}$$

Note that $\widehat{R}_1$ represents the proportion of household income spent on home loan mortgage and it equals to 43.75%. The s.e. estimate is

$$\begin{aligned}
s_{r1}^2 &= \frac{1}{n-1}\left(\sum_i y_i^2 - 2\widehat{R}_1\sum_i x_{1i}y_i + \widehat{R}_1^2\sum_i x_{1i}^2\right) \\
&= \frac{1}{7}\left(5439.616 - 2\times 0.4375 \times 12131.95 + 0.4375^2 \times 27796.23\right) \\
&= 20.64789 \\
\text{var}(\widehat{Y}_{r1}) &= N^2\left(1-\frac{n}{N}\right)\frac{s_{r1}^2}{n} = 832^2\left(1-\frac{8}{832}\right)\frac{20.64789}{8} = 1,769,441 \\
\text{se}(\widehat{Y}_{r1}) &= \sqrt{1,769,441} = 1330.204
\end{aligned}$$

When X_2 the household size is used, we have $\bar{X}_2 = 4.1$ and $\bar{x}_2 = 4.25$. The ratio estimators of the total amount of home loan payment is

$$\begin{aligned}
\widehat{R}_2 &= r_2 = \frac{\sum_i y_i}{\sum_i x_{2i}} = \frac{195.866}{34} = 5.7608 \\
\widehat{Y}_{r2} &= \widehat{R}_2 N\overline{X}_2 = 5.7608 \times 832 \times 4.1 = 19651.12
\end{aligned}$$

Note that $\widehat{R}_2$ represents the amount of home loan payment per individual and it equals 5.7608 thousand dollars. The s.e. estimate is

$$\begin{aligned}
s_{r2}^2 &= \frac{1}{n-1}\left(\sum_i y_i^2 - 2\widehat{R}_2\sum_i x_{2i}y_i + \widehat{R}_2^2\sum_i x_{2i}^2\right) \\
&= \frac{1}{7}\left(5439.616 - 2\times 5.7608 \times 804.711 + 5.7608^2 \times 156\right) = 192.1706 \\
\text{var}(\widehat{Y}_{r2}) &= N^2\left(1-\frac{n}{N}\right)\frac{s_{r2}^2}{n} = 832^2\left(1-\frac{8}{832}\right)\frac{192.1706}{8} = 16,468,254 \\
\text{se}(\widehat{Y}_{r2}) &= \sqrt{16,468,254} = 4058.11
\end{aligned}$$

(c) Since we have $\text{se}(\widehat{Y}_{r1}) < \text{se}(\widehat{Y}_{r2})$ and the correlation coefficients between Y the home loan payment and X_1 the household income is much more strongly and positively related ($\hat{\rho}_1 = 0.881$) than with X_2 the household size ($\hat{\rho}_2 = -0.322$), the auxiliary variable X_1 the household income is preferred.

(d) We may post-stratified the population of households with home loan mortgage on the household income resulting in more homogenous strata with respect to the amount of home loan payment. We may also draw cluster of living blocks first and then a random sample of households, resulting in a 2-stage cluster sampling

3. (a) The stratum weight $W_l = \frac{N_l}{N}$.

$$\begin{aligned}
N_1 &= N \times W_1 = 10,000 \times 0.55 = 5,500 \\
N_2 &= N \times W_2 = 10,000 \times 0.10 = 1,000 \\
N_3 &= N \times W_3 = 10,000 \times 0.35 = 3,500
\end{aligned}$$

(b) Estimate the average household income for households in the district using simple average and the standard error of the estimate are:

$$\begin{aligned}
\widehat{\overline{Y}}_{st,1} &= \frac{1}{L}\sum_l \bar{y}_l = \frac{1}{3}(\bar{y}_1+\bar{y}_2+\bar{y}_3) = \frac{1}{3}(2.7+3.5+7.2) = 4.467 \\
\text{var}(\widehat{\overline{Y}}_{st,1}) &= \frac{1}{L^2}\sum_l \text{var}(\bar{y}_l) = \frac{1}{L^2}\sum_l \left(1-\frac{n_l}{N_l}\right)\frac{S_l^2}{n_l} \\
&= \frac{1}{9}\left[\left(1-\frac{50}{5500}\right)\frac{4}{50} + \left(1-\frac{50}{1000}\right)\frac{11}{50} + \left(1-\frac{50}{3500}\right)\frac{85}{50}\right] = 0.2182
\end{aligned}$$

(c) Estimate the average household income for households in the district using SRS and stratified SRS with proportional allocation $W_l = \frac{n_l}{n}$:

$$\begin{aligned}
\widehat{\overline{Y}}_{st,2} &= \frac{1}{n}\sum_{l,i} y_{l,i} = \frac{1}{150}(n_l\bar{y}_1 + n_2\bar{y}_2 + n_3\bar{y}_3) = \frac{1}{3}(2.7+3.5+7.2) = 4.467 \\
\widehat{\overline{Y}}_{st,3} &= \sum_l W_l\bar{y}_l = \sum_l \frac{n_l}{n}\bar{y}_l = \frac{1}{3}(\bar{y}_1+\bar{y}_2+\bar{y}_3) = \frac{1}{3}(2.7+3.5+7.2) = 4.467
\end{aligned}$$

Note that the variance $\text{var}(\widehat{\overline{Y}}_{st,3}) = \text{var}(\widehat{\overline{Y}}_{st,1}) = 0.2182$ since both estimators $\widehat{\overline{Y}}_{st,1}$ and $\widehat{\overline{Y}}_{st,3}$ use stratified SRS estimators with the same weights $W_l = \frac{1}{3}$. However for $\widehat{\overline{Y}}_{st,2}$ when the sample is treated as a SRS, we have

$$\begin{aligned}
\sum_j y_{1j}^2 &= (n_1-1)s_1^2 + n_1\bar{y}_1^2 = 49(4) + 50(2.7^2) = 560.5 \\
\sum_j y_{2j}^2 &= (n_2-1)s_2^2 + n_2\bar{y}_2^2 = 49(11) + 50(3.5^2) = 1151.5 \\
\sum_j y_{3j}^2 &= (n_3-1)s_3^2 + n_3\bar{y}_3^2 = 49(85) + 50(7.2^2) = 6757.0 \\
\sum_{i,j} y_{ij}^2 &= \sum_j y_{1j}^2 + \sum_j y_{2j}^2 + \sum_j y_{3j}^2 = 560.5 + 1151.5 + 6757.0 = 8469 \\
s^2 &= \frac{1}{n-1}\left(\sum_{i,j} y_{ij}^2 - n\bar{y}^2\right) = \frac{1}{149}\left(8469 - 150\times 4.467^2\right) = 36.75391 \\
\text{var}(\widehat{\overline{Y}}_{st,2}) &= \left(1-\frac{n}{N}\right)\frac{s^2}{n} = \left(1-\frac{150}{10,000}\right)\frac{36.75391}{150} = 0.241351
\end{aligned}$$

In fact we have 3 SRSs of $n_i = 50$ each instead of a single SRS of $n = 150$. The assumption is NOT valid for $\text{var}(\widehat{\overline{Y}}_{st,2})$. We prefer the other two estimators $\widehat{\overline{Y}}_{st,1}$ and $\widehat{\overline{Y}}_{st,3}$ which are in fact equivalent.

(d) Estimate for the mean household income under stratified simple random sampling and an estimate of variance for this estimator are

$$\begin{aligned}
\widehat{\overline{Y}}_{st,1} &= \sum_l W_l\bar{y}_l = 0.55\times 2.7 + 0.10\times 3.5 + 0.35\times 7.2 = 4.355 \\
\text{var}(\widehat{\overline{Y}}_{st,1}) &= \sum_l W_l^2\,\text{var}(\bar{y}_l) = \sum_l W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_l^2}{n_l} \\
&= 0.55^2\left(1-\frac{50}{5500}\right)\frac{4}{50} + 0.10^2\left(1-\frac{50}{1000}\right)\frac{11}{50} + 0.35^2\left(1-\frac{50}{3500}\right)\frac{85}{50} = 0.2313
\end{aligned}$$

(e) Estimate for the difference in proportion of household which have household incomes under \$20,000 between households residing in public and rented private with a 95% confidence interval for the estimate are

$$\begin{aligned}
\widehat{P}_1 - \widehat{P}_2 &= p_1 - p_2 = 0.54 - 0.39 = 0.15 \\
\text{var}(\widehat{P}_1 - \widehat{P}_2) &= \frac{p_1(1-p_1)}{n_1 - 1} + \frac{p_2(1-p_2)}{n_2 - 1} = \frac{0.54 \times 0.46}{49} + \frac{0.39 \times 0.61}{49} = 0.0099 \\
\text{se}(\widehat{P}_1 - \widehat{P}_2) &= \sqrt{0.0099} = 0.09962 \\
95\% \text{ C.I. for} \widehat{P}_1 - \widehat{P}_2 &= 0.15 \pm 1.96 \times 0.09962 = (-0.045,\ 0.345)
\end{aligned}$$

Since it includes 0, the difference $\widehat{P}_1 - \widehat{P}_2$ is not significant at 0.05 level.

(f) For Neyman allocation

$$\begin{aligned}
\sum_{l=1}^{L} N_l s_l &= N_1 s_1 + N_2 s_2 + N_3 s_3 = 5500\sqrt{4} + 1000\sqrt{11} + 3500\sqrt{85} = 46,585.03 \\
n_1 &= n w_1 = n\left(\frac{N_1 s_1}{\sum_{l=1}^{L} N_l^2 s_l}\right) = 300\left[\frac{5500\sqrt{4}}{46,585.03}\right] = 70.84 = 71 \\
n_2 &= n w_2 = 300\left[\frac{1000\sqrt{11}}{46,585.03}\right] = 21.36 = 21 \\
n_3 &= n w_3 = 300\left[\frac{3500\sqrt{85}}{46,585.03}\right] = 207.8 = 208
\end{aligned}$$

4. (a) Estimate of the average number of persons per family and the variance estimate are

$$\begin{aligned}
\widehat{\overline{X}} &= \frac{1}{n}\sum_i x_i = \frac{111}{30} = 3.7 \\
s_x^2 &= \frac{1}{n-1}\left(\sum_i x_i^2 - n\bar{y}^2\right) = \frac{1}{29}(477 - 30 \times 3.7^2) = 2.286207 \\
\text{var}\left(\widehat{\overline{X}}\right) &= \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = \left(1 - \frac{30}{660}\right)\frac{2.286207}{30} = 0.0727
\end{aligned}$$

(b) Estimate of the proportion of family income that is spent on food

$$\widehat{R}_1 = r_1 = \frac{\sum_i z_i}{\sum_i y_i} = \frac{848.2}{2195} = 0.3864$$

Estimate of the average expenditure on food in thousand per family and the variance of this estimator are

$$\begin{aligned}
\widehat{\overline{Z}}_r &= r_1 \times \overline{Y} = 0.3864 \times 72 = 27.8225 \\
s_r^2 &= \frac{1}{n-1}\left(\sum_i z_i^2 - 2\widehat{R}_1 \sum_i z_i y_i + \widehat{R}_1^2 \sum_i y_i^2\right) \\
&= \frac{1}{29}\left[27,060.26 - 2 \times \frac{848.2}{2,195} \times 62,771.4 + \left(\frac{848.2}{2,195}\right)^2 \times 164,305\right] = 106.2796 \\
\text{var}(\widehat{\overline{Z}}_r) &= \left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} = \left(1 - \frac{30}{660}\right)\frac{106.2796}{30} = 3.381624
\end{aligned}$$

(c) The estimate of the average expenditure on food in thousand per individual using the usual ratio estimator is

$$\widehat{R}_2 = \frac{\overline{z}}{\overline{x}} = \frac{848.2}{111} = 7.64$$

Another estimator using the results in part (a) and (b) is

$$\widehat{R}_3 = \frac{\widehat{\overline{Z}}_r}{\overline{x}} = \frac{\overline{X}\,\overline{z}}{\overline{x}\,\overline{y}} = \frac{27.8225}{3.7} = 7.52$$

Both are ratio estimates. The first one makes use of sample information only whereas the second one use the additional information of the family income from certain source. The first estimator is preferred as its variance estimate can be obtained. The second estimate is more complicated and its variance estimate is difficult to obtain.