STAT 3014/3914

Semester 2	Applied Statistics	2015

Solution to Tutorial 13

1. Note that y_i is the total mileage for branch i.

   (a) 1-stage cluster sample

$$\begin{aligned}
\text{Cluster} &\quad - \quad \text{branches } (N = 12;\ n = 4) \\
\text{Element} &\quad - \quad \text{cars } (M = 810;\ m = 240) \\
\text{Population mean no. of cars per branch } \bar{M} &\quad = \quad \frac{M}{N} = \frac{810}{12} = 67.5 \\
\text{Sample mean mileage per branch } \bar{y} &\quad = \quad \frac{\sum_{i=1}^{n} y_i}{n} = \frac{6340.9}{4} = 1585.225 \\
\text{Sample mean mileage per car } r &\quad = \quad \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i} = \frac{6340.9}{240} = 26.4204.
\end{aligned}$$

The sampling method is a single-stage cluster sampling and the quantity to be estimated is R, the average mileage per car.

The ordinary estimate of mean mileage per car is

$$\widehat{R}_{c1} = \frac{\bar{y}}{\bar{M}} = \frac{1585.225}{67.5} = 23.4848$$

with estimated variance

$$\text{var}(\widehat{R}_{c1}) = \frac{1}{\bar{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = \frac{1}{67.5^2}\left(1 - \frac{4}{12}\right)\frac{1083221.642}{4} = 39.624.$$

The ratio estimate of mean mileage per car is

$$\widehat{R}_{c1,r} = r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i} = \frac{6340.9}{240} = 26.4204$$

with estimated variance

$$
\begin{aligned}
s_r^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rM_i)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - 2r\sum_{i=1}^{n} y_i M_i + r^2 \sum_{i=1}^{n} M_i^2\right) \\
&= \frac{13301418.13 - 2\cdot 26.4204 \cdot 496751 + 26.4204^2 18600}{3} = \frac{36195.885}{3} \\
\mathrm{var}(\widehat{R}_{c1,r}) &= \frac{1}{\bar{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} \\
&= \frac{1}{67.5^2}\left(1 - \frac{4}{12}\right)\frac{1}{4}\frac{36195.883}{3} = 0.441.
\end{aligned}
$$

Note: The ordinary estimate has a much larger variance (39.624) than the ratio estimate (0.441). This is due to the great variability in cluster sizes ($M_i = 60, 110, 20, 50$ for the selected clusters).

(b) 2-stage cluster sample

$$
\begin{aligned}
\text{Cluster} &\quad - \quad \text{branches } (N = 12;\ n = 4) \\
\text{Element} &\quad - \quad \text{cars } (M = 810;\ m = 240) \\
\text{Mean no. of car per branch } \bar{M} &= \frac{M}{N} = \frac{810}{12} = 67.5 \\
\text{Estimated sample mean mileage per branch } \hat{\bar{y}} &= \frac{\sum_{i=1}^{n}\hat{y}_i}{n} = \frac{6,521.58}{4} = 1,630.395 \\
\text{Estimated sample mean mileage per car } \hat{r} &= \frac{\sum_{i=1}^{n}\hat{y}_i}{\sum_{i=1}^{n} M_i} = \frac{6,521.58}{240} = 27.17325.
\end{aligned}
$$

Branch	M_i	y_i	$M_i\hat{r}$	$(y_i - M_i\hat{r})$	$(y_i - M_i\hat{r})^2$
1	60	1459.2	1585.225	-126.025	15882.301
2	110	3036.0	2906.246	129.754	16836.144
3	20	568.2	528.408	39.792	1583.377
4	50	1277.5	1321.021	-43.521	1894.063
Sum	240	6340.9			36195.885

Variance due to estimated $\hat{y}_i$ is

$$
\begin{aligned}
\frac{N}{nM^2}\sum_{i=1}^{n}\mathrm{var}(\hat{y}_i) &= \frac{N}{nM^2}\sum_{i=1}^{n} M_i^2\left(1 - \frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i} \\
&= \frac{12}{4\cdot 810^2}\left[60^2\left(1 - \frac{10}{60}\right)\frac{4.90684}{10} + \cdots + 50^2\left(1 - \frac{10}{50}\right)\frac{13.42161}{10}\right] \\
&= \frac{12}{4\cdot 810^2}(1,472.052 + 13,811.743 + 121.8964 + 2,684.322) \\
&= \frac{54,270.04}{810^2}.
\end{aligned}
$$

We also have

$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} \hat{y}_i^2 - n\bar{\hat{y}}^2\right) = \frac{1}{3}[14,161,800.33 - 4(1,630.395)^2] = 1,176,349.635.$$

$$\begin{aligned} s_r^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} \hat{y}_i^2 - 2\hat{r}\sum_{i=1}^{n} \hat{y}_i M_i + \hat{r}^2 \sum_{i=1}^{n} M_i^2\right) \\ &= \frac{1}{3}(14,161,800.33 - 2 \times 27.173 \times 512,461.2 + 27.173^2 \times 18,600) \\ &= 15,099.44. \end{aligned}$$

The ordinary estimate of the average mileage per car R is

$$\widehat{R}_{c2} = \frac{\bar{\hat{y}}}{\bar{M}} = \frac{1,630.395}{67.5} = 24.154$$

with estimated variance:

$$\text{var}(\widehat{R}_{c2}) = \frac{1}{\bar{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = \frac{12^2}{810^2}\left(1 - \frac{4}{12}\right)\frac{1,176,349.635}{4} = \frac{28,232,391.24}{810^2}.$$

Thus

$$\begin{aligned} \text{var}(\widehat{R}_{c2}) &= \text{var}(\widehat{R}_{c1}) + \frac{N}{nM^2}\sum_{i=1}^{n} \text{var}(\hat{y}_i) \\ &= \frac{28,232,391.24 + 54,270.04}{810^2} = 43.0306 + 0.0827 = 43.1133. \end{aligned}$$

The ratio estimate of the average mileage per car R is

$$\widehat{R}_{c2,r} = \hat{r} = \frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} M_i} = \frac{6,521.58}{240} = 27.173$$

with estimated variance

$$\text{var}(\widehat{R}_{c1,r}) = \frac{1}{\bar{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} = \frac{12^2}{810^2}\left(1 - \frac{4}{12}\right)\frac{15,099.44}{4} = \frac{362,386.544}{810^2}.$$

Thus

$$\begin{aligned} \text{var}(\widehat{R}_{c2,r}) &= \text{var}(\widehat{R}_{c1,r}) + \frac{N}{nM^2}\sum_{i=1}^{n} \text{var}(\hat{y}_i) \\ &= \frac{362,386.544 + 54,270.04}{810^2} = 0.5523 + 0.0827 = 0.635. \end{aligned}$$

Note:

1. The ordinary estimate has a much larger variance (43.1133) than the ratio estimate (0.635). This is due to the great variability in cluster sizes ($N_i = 60, 110, 20, 50$ for the selected clusters).
2. Both estimates of ordinary and ratio in 2-stage sampling have larger variance than the corresponding estimates in singe-stage sampling due to the increase in variability in estimating $\hat{y}_i$ in the 2-stage sampling. However the increase (0.0837) is not great since the subsample sizes of 10 in each selected cluster are not too small.

2. 2-stage cluster sample

Plant	M_i	m_i	$\overline{\overline{y}}_i$	$\hat{y}_i = M_i\overline{\overline{y}}_i$	s^2_{yi}
1	50	10	5.40	270.00	11.38
2	65	13	4.00	260.00	10.67
3	45	9	5.67	255.15	16.75
4	48	10	4.80	230.40	13.29
5	52	10	4.30	223.60	11.12
6	58	12	3.83	222.14	14.88
7	42	8	5.00	210.00	5.14
8	66	13	3.85	254.10	4.31
9	40	8	4.88	195.20	6.13
10	56	11	5.00	280.00	11.80

$$\sum_i \hat{y}_i = 2400.59,\ \sum_i M_i = 522,\ \sum_i \hat{y}_i^2 = 583,198.6721,\ \sum_i \hat{y}_i M_i = 126,530.87,\ \sum_i M_i^2 = 27978$$

$$\begin{aligned}
\text{Cluster} \quad &- \quad \text{plant } (N = 90;\ n = 10)\\
\text{Element} \quad &- \quad \text{machine } (M = 4,500;\ m = 522)\\
\text{Mean no. of machine per plant } \overline{M} \quad &= \quad \frac{M}{N} = \frac{4,500}{90} = 50\\
\text{Mean downtime per plant } \overline{y} \quad &= \quad \frac{\sum_{i\in\mathcal{S}} \hat{y}_i}{n} = \frac{2,400.59}{10} = 240.059\\
\text{Mean downtime per machine } \overline{\overline{y}} \quad &= \quad \frac{\sum_{i\in\mathcal{S}} \hat{y}_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{2,400.59}{522} = 4.5988.
\end{aligned}$$

(a) The ordinary estimate of the average downtime per machine $\overline{\overline{Y}}$ is

$$\widehat{\overline{\overline{Y}}}_{c2} = \frac{\overline{\hat{y}}}{\overline{M}} = \frac{240.059}{50} = 4.80118$$

with estimated variance for stage 1 as

$$\text{var}(\widehat{\overline{\overline{Y}}}_{c1}) = \frac{1}{\overline{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = \frac{1}{50^2}\left(1 - \frac{10}{90}\right)\frac{768.38}{10} = 0.027320246.$$

where

$$s_y^2 = \frac{\sum_{i\in\mathcal{S}}(\hat{y}_i - \overline{\hat{y}})^2}{n-1} = \frac{\sum_{i\in\mathcal{S}}\hat{y}_i^2 - n\overline{\hat{y}}^2}{n-1} = \frac{1}{3}[583,198.6721 - 10(240.059)^2] = 768.38$$

Variance due to estimated $\hat{y}_i$ is

$$\begin{aligned}
\frac{N}{nM^2}\sum_{i\in\mathcal{S}}\mathrm{var}(\widehat{y}_i) &= \frac{N}{nM^2}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i} \\
&= \frac{90}{10\cdot 4500^2}\left[50^2\left(1-\frac{10}{50}\right)\frac{11.38}{10} + 65^2\left(1-\frac{13}{65}\right)\frac{10.67}{13} + \cdots + \right. \\
&\qquad \left. 56^2\left(1-\frac{11}{56}\right)\frac{11.80}{11}\right] \\
&= \frac{90}{10\cdot 4500^2}(26,285.475) = 0.01168243.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{var}(\widehat{\overline{\overline{Y}}}_{c2}) &= \mathrm{var}(\widehat{\overline{\overline{Y}}}_{c1}) + \frac{N}{nM^2}\sum_{i\in\mathcal{S}}\mathrm{var}(\widehat{y}_i) \\
&= 0.02732025 + 0.01168243 = 0.03900268 \\
\text{Error bound}(\widehat{\overline{\overline{Y}}}_{c2}) &= 1.96\ \sqrt{0.03900268} = 1.96(0.19749096) = 0.38708228
\end{aligned}$$

(b) The ratio estimate of the average downtime per machine $\overline{\overline{Y}}$ is

$$\widehat{\overline{\overline{Y}}}_{c2,r} = \frac{\sum_{i\in\mathcal{S}}\hat{y}_i}{\sum_{i\in\mathcal{S}}M_i} = \frac{2,400.59}{522} = 4.5988$$

with estimated variance for stage 1:

$$\mathrm{var}(\widehat{\overline{\overline{Y}}}_{c1,r}) = \frac{1}{\overline{M}^2}\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} = \frac{1}{50^2}\left(1-\frac{10}{90}\right)\frac{1,236.01328}{10} = 0.043947138.$$

where

$$\begin{aligned}
s_r^2 &= \frac{\sum_{i\in\mathcal{S}}(\hat{y}_i - M_i\overline{\overline{\hat{y}}})^2}{n-1} = \frac{\sum_{i\in\mathcal{S}}\hat{y}_i^2 - 2\overline{\overline{\hat{y}}}\sum_i \hat{y}_i M_i + \overline{\overline{\hat{y}}}^2\sum_i M_i^2}{n-1} \\
&= \frac{1}{9}(583,198.6721 - 2\times 4.5988\times 126,530.87 + 4.5988^2\times 27,978) = 1,236.01328
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{var}(\widehat{\overline{\overline{Y}}}_{c2,r}) &= \mathrm{var}(\widehat{\overline{\overline{Y}}}_{c1,r}) + \frac{N}{nM^2}\sum_{i\in\mathcal{S}}\mathrm{var}(\widehat{y}_i) \\
&= 0.043947138 + 0.01168243 = 0.05562957 \\
\text{Error bound}(\widehat{\overline{\overline{Y}}}_{c2,r}) &= 1.96\ \sqrt{0.05562957} = 1.96(0.23585922) = 0.46228407
\end{aligned}$$

Note that $s_r^2 > s_y^2$ because $\hat{\rho} = 0.54316681 < 0.74687872 = \frac{s_x \overline{y}}{2 s_y \overline{x}}$.

3. (a) HH estimator:

   (i) Six-digit random numbers are generated, ignoring 000000 and any numbers greater than 186030. If the list are 001052, 185953, 000600, 000987 say, the selected hospitals are 2, 1158, 1 and 2 such that hospital 2 appears twice in the sample.

   (ii) We have $n = 4$, $N = 1,158$, $\sum_{i=1}^{n} \frac{y_i}{p_i} = 563,602$, $\sum_{i=1}^{n} \left(\frac{y_i}{p_i}\right)^2 = 79,470,194,284$, $\sum_{i=1}^{n} \frac{y_i'}{p_i} = 399.62$ and $\sum_{i=1}^{n} \left(\frac{y_i'}{p_i}\right)^2 = 91,336.32$

$$
\begin{aligned}
\widehat{Y}_{HH} &= \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} = \frac{1}{4}\left(\frac{350}{0.0024}+\frac{1,100}{0.0081}+\frac{500}{0.0036}+\frac{350}{0.0024}\right) = 140,900.5 \\
\widehat{\overline{Y}}_{HH} &= \frac{1}{n \times N}\sum_{i=1}^{n}\frac{y_i}{p_i} \\
&= \frac{1}{4 \times 1,158}\left(\frac{350}{0.0024}+\frac{1,100}{0.0081}+\frac{500}{0.0036}+\frac{350}{0.0024}\right) = 121.6757 \text{ (\$000)} \\
s_{y/p}^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 - n\overline{\frac{y}{p}}^2\right] \\
&= \frac{1}{3}(79,470,194,284 - 4 \times 140,900.5^2) = 19,463,561 \\
\mathrm{var}(\widehat{\overline{Y}}_{HH}) &= \frac{1}{N^2}\frac{s_{y/p}^2}{n} = \frac{1}{1,158^2}\frac{19,463,561}{4} = 3.628651 \\
\mathrm{se}(\widehat{\overline{Y}}_{HH}) &= \sqrt{3.628651} = 1.904902
\end{aligned}
$$

$$
\begin{aligned}
\widehat{P}_{HH} &= \frac{1}{nN}\sum_{i=1}^{n}\frac{y_i'}{p_i} \\
&= \frac{1}{4 \times 1,158}\left(\frac{0}{0.0024}+\frac{1}{0.0081}+\frac{1}{0.0036}+\frac{0}{0.0024}\right) = 0.086274 \\
s_{y'/p}^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\frac{y_i'}{p_i}\right)^2 - n\overline{\frac{y'}{p}}^2\right] \\
&= \frac{1}{3}(91,336.32 - 4 \times 99.905^2) = 17,137.43 \\
\mathrm{var}(\widehat{\overline{Y}}_{HH}) &= \frac{1}{N^2}\frac{s_{y'/p}^2}{n} = \frac{1}{1,158^2}\frac{17,137.43}{4} = 0.003195 \\
\mathrm{se}(\widehat{\overline{Y}}_{HH}) &= \sqrt{0.003195} = 0.056524
\end{aligned}
$$

The total estimates are respectively $(1,158)(121.6757) = 140,900.5$ ($000) and $(1,158)(0.086274) = 100$.

(b) With IPPS and $n = 3$:

(i) The HT estimate is:

$$\widehat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i} = \frac{1}{1158}\left(\frac{500}{\frac{15}{1378}} + \frac{350}{\frac{5}{689}} + \frac{1100}{\frac{50}{2067}}\right) = 120.5849 \text{ (\$000)}$$

The total estimate of hospital purchases and count for product Y is $(1,158)(120.5849) = 139,637.3$ ($000) It is hard to find their s.e. unless we know π_{ij}. *If we can assume the usual the draw-by-draw p_i for sampling with replacement*, the second order inclusion probabilities are

$$\begin{aligned}
\pi_1 &= 1-(1-p_1)^3 = \frac{15}{1378} \Rightarrow p_1 = 1-\left(1-\frac{15}{1378}\right)^{1/3} = 0.003641693 \\
\pi_2 &= 1-(1-p_2)^3 = \frac{5}{689} \Rightarrow p_2 = 1-\left(1-\frac{5}{689}\right)^{1/3} = 0.002424840 \\
\pi_{1158} &= 1-(1-p_{1158})^3 = \frac{50}{2067} \Rightarrow p_{1158} = 1-\left(1-\frac{50}{2067}\right)^{1/3} = 0.008129119 \\
\pi_{1,2} &= \pi_1+\pi_2-[1-(1-p_1-p_2)^3] \\
&= \underbrace{\frac{15}{1378}+\frac{5}{689}}_{12,\,1\bar{2},\,12,\,\bar{1}2} - \underbrace{[1-(1-0.003641693-0.002424840)^3]}_{12,\,1\bar{2},\,\bar{1}2} = 0.00005282242 \\
\pi_{1,1158} &= \pi_1+\pi_{1158}-[1-(1-p_1-p_{1158})^3] \\
&= \frac{15}{1378}+\frac{50}{2067}-[1-(1-0.003641693-0.008129119)^3] = 0.0001765771 \\
\pi_{2,1158} &= \pi_2+\pi_{1158}-[1-(1-p_2-p_{1158})^3] \\
&= \frac{5}{689}+\frac{50}{2067}-[1-(1-0.002424840-0.008129119)^3] = 0.0001176468
\end{aligned}$$

Note that the draw-by-draw sampling is a required assumption when calculating the second order inclusion probabilities. We have $\pi_1 = \frac{15}{1378} = 0.01088534$, $\pi_2 = \frac{5}{689} = 0.007256894$, $\pi_{1158} = \frac{50}{2067} = 0.024189647$.

$$
\begin{aligned}
\mathrm{var}(\widehat{Y}_{HT,1}) &= \frac{1}{N^2}\left(\sum_{i\in\mathcal{S}}\frac{1-\pi_i}{\pi_i^2}\; y_i^2 + 2\sum_{i<j}\sum\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}\; y_iy_j\right)\\
&= \frac{1}{1158^2}\left[\frac{1-0.01088534}{0.01088534^2}500^2+\frac{1-0.007256894}{0.007256894^2}350^2+\frac{1-0.024189647}{0.024189647^2}1100^2+\right.\\
&\quad 2\left(\frac{0.00005282242-0.01088534\times 0.007256894}{0.01088534\times 0.007256894\times 0.00005282242}\;500\times 350+\right.\\
&\quad \frac{0.0001765771-0.01088534\times 0.024189647}{0.01088534\times 0.024189647\times 0.0001765771}\;500\times 1100+\\
&\quad \left.\left.\frac{0.0001176468-0.007256894\times 0.024189647}{0.007256894\times 0.024189647\times 0.0001176468}\;350\times 1100\right)\right] = 6.080117\\
\mathrm{se}(\widehat{\overline{Y}}_{HT,1}) &= \sqrt{6.080117} = 2.465789
\end{aligned}
$$

(ii) With systematic IPPS sampling and $n=3$, $\widehat{\overline{Y}}_{sys,pps} = 120.269$, same as (i) but the se estimate is

$$
\begin{aligned}
\mathrm{var}(\widehat{\overline{Y}}_{sys,pps}) &\simeq \frac{1}{N^2}\sum_{i\in\mathcal{S}}\left(1-\frac{n-1}{n}\pi_i\right)\left(\frac{y_i}{\pi_i}-\frac{\widehat{Y}_{sys}}{n}\right)^2\\
&= \frac{1}{1,158^2}\left[\left(1-\frac{2}{3}\,0.010885\right)\left(\frac{500}{0.010885}-\frac{139,637.3}{3}\right)^2+\right.\\
&\quad \left(1-\frac{2}{3}\,0.007257\right)\left(\frac{350}{0.007257}-\frac{139,637.3}{3}\right)^2+\\
&\quad \left.\left(1-\frac{2}{3}\,0.024190\right)\left(\frac{1100}{0.024190}-\frac{139,637.3}{3}\right)^2\right]\\
&= 3.225613\\
\mathrm{se}(\widehat{\overline{Y}}_{sys,pps}) &= \sqrt{3.225613} = 1.795999
\end{aligned}
$$

4. The second order inclusion probabilities are

(i,j)		π_{ij}	sum	(i,j)		π_{ij}	Sum
(1,2)	JB,JS	$\frac{1}{2}\cdot 1\cdot\frac{1}{2}$		(4,9)	AJ,DE	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}$	
(2,3)	MB,JS	$\frac{1}{2}\cdot 1\cdot\frac{1}{2}$	$\frac{1}{2}$	(5,9)	PJ,DE	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}$	$\frac{1}{6}$
(4,6)	AJ,JC	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{3}$		(6,9)	JC,DE	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{3}$	
(5,6)	PJ,JC	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{3}$		(7,9)	SC,DE	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{3}$	
(4,7)	AJ,SC	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{3}$		(8,9)	MC,DE	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{3}$	$\frac{1}{6}$
(5,7)	PJ,SC	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{3}$					
(4,8)	AJ,MC	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{3}$					
(5,8)	PJ,MC	$\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{3}$	$\frac{1}{6}$				

Note that this sampling scheme defines a set of π_i and π_{ij} but it is not an inclusion probability proportional to size (IPPS) sampling as there is no X variable which defines $\pi_i = \dfrac{nx_i}{X}$ even though π_i varies across individuals.

$$
\begin{aligned}
\widehat{\overline{Y}}_{HT} &= \frac{1}{N} \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} = \frac{1}{9}\left(\frac{25}{1/6} + \frac{60}{1/3}\right) = 36.67 \\
\operatorname{var}(\widehat{\overline{Y}}_{HT,1}) &= (1-\pi_1)\frac{y_1^2}{\pi_1^2} + (1-\pi_2)\frac{y_2^2}{\pi_2^2} + 2\left(\frac{\pi_{12} - \pi_1\pi_2}{\pi_{12}}\frac{y_1 y_2}{\pi_1\pi_2}\right) \\
&= \left(1 - \frac{1}{6}\right)\frac{25^2}{(\frac{1}{6})^2} + \left(1 - \frac{1}{3}\right)\frac{60^2}{(\frac{1}{3})^2} + 2\left(1 - \frac{\frac{1}{6}\frac{1}{3}}{\frac{1}{12}}\right)\frac{25(60)}{\frac{1}{6}\frac{1}{3}} = 594.4444 \\
\operatorname{se}(\widehat{\overline{Y}}_{HT,1}) &= \sqrt{594.4444} = 24.38 \\
\operatorname{var}(\widehat{\overline{Y}}_{HT,2}) &= = \frac{1}{N^2}\left(\frac{\pi_1\pi_2 - \pi_{1,2}}{\pi_{1,2}}\right)\left(\frac{y_1}{\pi_2} - \frac{y_2}{\pi_2}\right)^2 = \frac{1}{9^2}\left(\frac{\frac{1}{6}\frac{1}{3} - \frac{1}{12}}{\frac{1}{12}}\right)\left(\frac{25}{1/6} - \frac{60}{1/3}\right)^2 = -\text{ve}
\end{aligned}
$$

Formula one can be applied to any sampling scheme which defines a set of π_i and π_{ij} but formula two can only be applied to sampling schemes without replacement.

Extra exercise

1. 2-stage cluster sample. We have $N = 30$ and $n = 3$.

i	M_i	m_i	**Sample data** y_{ij}	**Sample mean** $\overline{\overline{y}}_i$	**Sample var.** s_{yi}^2	$\hat{y}_i = M_i\overline{\overline{y}}_i$
1	10	4	1, 3, 3, 4	2.75	1.5833	27.50
2	12	4	3, 4, 0, 1	2.00	3.3333	24.00
3	9	4	4, 2, 0, 1	1.75	2.1875	15.75
Total	31	12			67.25	

$$\sum_i M_i = 31; \ \sum_i M_i^2 = 325; \ \sum_i \hat{y}_i = 67.25; \ \sum_i \hat{y}_i^2 = 1,580.3125; \ \sum_i \hat{y}_i M_i = 704.75$$

Additional variance due to those households of size ≥ 4:

$$
\begin{aligned}
&\frac{N}{n}\sum_i M_i^2\left(1 - \frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i} = \frac{N}{n}\sum_i M_i^2\left(1 - \frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i} \\
= &\frac{30}{3}\left[10^2\left(1 - \frac{4}{10}\right)\frac{1.5833}{4} + 12^2\left(1 - \frac{4}{12}\right)\frac{3.3333}{4} + 9^2\left(1 - \frac{4}{9}\right)\frac{2.1875}{4}\right] \\
= &10 \times 128.358075 = 1,283.58075
\end{aligned}
$$

Ratio estimator of total:

$$\begin{aligned}
\widehat{Y}_{c2,r} &= M \times \hat{\bar{\bar{y}}} = M \times \frac{\sum_{i\in\mathcal{S}} \hat{y}_i}{\sum_{i\in\mathcal{S}} M_i} = 315 \times \frac{67.25}{31} = 315 \times 2.169354839 = 683.3467743 \\
s_r^2 &= \frac{1}{n-1}\left(\sum_{i\in\mathcal{S}} \hat{y}_i^2 - 2\hat{\bar{\bar{y}}} \sum_{i\in\mathcal{S}} \hat{y}_i M_i + \hat{\bar{\bar{y}}}^2 \sum_{i\in\mathcal{S}} M_i^2\right) \\
&= \frac{1}{2}(1,580.3125 - 2 \cdot 2.169354839 \cdot 704.75 + 2.169354839^2 \cdot 325) \\
&= 26.04474505 \\
\mathrm{var}(\widehat{Y}_{c1,r}) &= N^2\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} = 30^2\left(1-\frac{3}{30}\right)\frac{26.04474505}{3} = 7,032.081163 \\
\mathrm{var}(\widehat{Y}_{c2,r}) &= \mathrm{var}(\widehat{Y}_{c1,r}) + \text{ Add. variance} \\
&= 7,032.081163 + 1,283.5807 = 8,315.661913 \\
\mathrm{var}(\widehat{Y}_{c2,r}) &= \sqrt{8,315.661913} = 91.1902512
\end{aligned}$$

Naive estimator of total:

$$\begin{aligned}
\widehat{Y}_{c1} &= N \times \hat{\bar{y}} = N \times \frac{\sum_i y_i}{n} = 30 \times \frac{67.25}{3} = 672.5 \\
s_y^2 &= \frac{1}{n-1}\left(\sum_{i\in\mathcal{S}} y_i^2 - n\hat{\bar{y}}^2\right) = \frac{1}{2}(1,580.3125 - 3 \cdot 22.4167^2) = 36.3958334 \\
\mathrm{var}(\widehat{Y}_{c2}) &= N^2\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} + \frac{N}{n}\sum_i M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i} \\
&= 30^2\left(1-\frac{3}{30}\right)\frac{36.3958334}{3} + 1,283.5807 \\
&= 9,826.875017 + 1,283.58075 = 11,110.45577 \\
\mathrm{var}(\widehat{Y}_{c2}) &= \sqrt{11,110.45577} = 105.4061467
\end{aligned}$$

The s.e. of ordinary estimate is slightly larger but the two s.e. are quite close to each other because the cluster total $\hat{y}_i$ is not highly correlated to the cluster size M_i and also the cluster sizes M_i are all close to 10. We still prefer ratio estimator to ordinary estimator as it uses the information of cluster size M_i. The two estimators will be the same if the cluster sizes M_i are all equal.

2. We have $\overline{M} = \dfrac{M}{N} = \dfrac{426}{41} = 10.3902439$ and $\bar{y} = \dfrac{\sum\limits_i y_i}{n} = \dfrac{279}{4} = 69.75$.

   (a) 1-stage cluster sampling:

(i) Ordinary estimator for mean per element:

$$\begin{aligned}
\widehat{\overline{\overline{Y}}}_{c1} &= \frac{1}{\overline{\overline{M}}} \times \hat{\bar{y}} = \frac{69.75}{10.3902439} = 6.713028169 \\
s_y^2 &= \frac{1}{n-1}(\sum_{i\in\mathcal{S}} y_i^2 - n\hat{\bar{y}}^2) = \frac{1}{3}(20,009 - 4\cdot 69.75^2) = 182.91\dot{6} \\
\mathrm{var}(\widehat{\overline{\overline{Y}}}_{c1}) &= \frac{1}{\overline{\overline{M}}^2}\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} = \frac{1}{10.3902439^2}\left(1-\frac{4}{41}\right)\frac{182.91\dot{6}}{4} \\
&= 0.382260716 \\
\mathrm{se}(\widehat{\overline{\overline{Y}}}_{c1}) &= \sqrt{0.382260716} = 0.618272363
\end{aligned}$$

(ii) Ordinary estimator is preferred because the cluster size differs only slightly and it is easier to compute. The ratio and ordinary estimates should be close as the cluster sizes are similar.

(b) 2-stage cluster sampling:

(i) Calculation:

i	m_i	Mark y_{ij}	$\hat{y}_i = M_i\overline{\overline{y}}_i = M_i \times \frac{\sum_j y_{ij}}{m_i}$	M_i	s_{yi}^2
1	5	7, 6, 4, 8, 10	$10 \times 35/5 = 70$	10	5
2	6	8, 9, 10, 9, 7, 7	$11 \times 50/6 = 91.\dot{6}$	11	1.47
3	5	7, 6, 4, 8, 9	$9 \times 34/5 = 61.2$	9	3.7
4	5	7, 8, 9, 7, 2	$10 \times 33/5 = 66$	10	7.3
5	5	6, 7, 4, 5, 4	$9 \times 26/5 = 46.8$	9	1.7
6	6	5, 2, 5, 4, 1, 0	$11 \times 17/6 = 31.1\dot{6}$	11	4.57
7	6	4, 6, 4, 3, 9, 2	$11 \times 28/6 = 51.\dot{3}$	11	6.27
8	5	8, 7, 5, 7, 9	$9 \times 36/5 = 64.8$	9	2.2

We have

$$\sum_i M_i = 80;\ \sum_i M_i^2 = 806;\ \sum_i \hat{y}_i = 482.9\dot{6};\ \sum_i \hat{y}_i^2 = 31,399.97;\ \sum_i \hat{y}_i M_i = 4,831.0\dot{3}$$

(ii) Ordinary estimator for mean per element:

$$\begin{aligned}
\hat{\bar{y}} &= \frac{\sum_i \hat{y}_i}{n} = \frac{482.9\dot{6}}{8} = 60.3708\dot{3} \\
\widehat{\overline{\overline{Y}}}_{c2} &= \frac{1}{\overline{\overline{M}}}\hat{\bar{y}} = \frac{60.3708\dot{3}}{10.3902439} \\
s_y^2 &= \frac{1}{n-1}(\sum_i \hat{y}_i^2 - n\hat{\bar{y}}^2) = \frac{1}{7}(31,399.97 - 8 \times 60.3708\dot{3}^2) = 320.5249714
\end{aligned}$$

$$
\begin{aligned}
\text{Add. var.} &= \frac{1}{nN\overline{\overline{M}}}\sum_i M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i} \simeq \frac{1}{nN\overline{\overline{M}}}\left(1-\frac{1}{2}\right)\sum_i M_i^2\frac{s_{yi}^2}{m_i} \\
&= \frac{1}{2\times 8\times 41\times 10.3902439^2}\left(10^2\frac{5}{5}+11^2\frac{1.47}{6}+9^2\frac{3.7}{5}+10^2\frac{7.3}{5}\right. \\
&\quad \left.+9^2\frac{1.7}{5}+11^2\frac{4.57}{6}+11^2\frac{6.27}{6}+9^2\frac{2.2}{5}\right) = 0.00871748824 \\
\text{var}(\hat{\overline{\overline{Y}}}_{c2}) &= \text{var}(\hat{\overline{\overline{Y}}}_{c1}) + \text{Add. var. due to } \hat{y}_i \\
&= \frac{1}{\overline{M}^2}\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} + \frac{1}{nN\overline{M}^2}\sum_i M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i} \\
&= \frac{1}{10.3902439^2}\left(1-\frac{8}{41}\right)\frac{320.5249714}{8} + 0.00871748824 \\
&= 0.298710495 + 0.00871748824 = 0.307427983 \\
\text{se}(\hat{\overline{\overline{Y}}}_{c2}) &= \sqrt{0.307427983} = 0.554461886
\end{aligned}
$$

(iii) The first sample of 4 classes consist mainly classes of high marks. As a result, $\hat{\overline{\overline{Y}}}_{c1}$ will overestimate the true mark. Also the s.e. $\text{se}(\hat{\overline{\overline{Y}}}_{c1})$ based on mainly classes of large total marks will underestimate the true s.e.. This is the result of sampling error.

(iv) Since the additional 4 classes are mainly classes of low marks, this shows great variability of marks across classes. The two-stage cluster sampling is preferred as more classes can be selected from the classes with more variability in class totals.

3. (a) (i) When the variability in cluster size M_i is large, 2-stage cluster sampling is preferred as we can subsample from those large cluster. As a result, the total sample size is easier to control.

(ii) When the variability of Y, the variable of interest within the cluster is relatively less than that between clusters, the 2-stage cluster sampling is preferred as it enables the selection of more clusters for a given total sample size of elements. The selection of more cluster is necessary as the variability between clusters is high.

(b) We have $n = 6$, $N = 48$, $M = 5,200$ and $\overline{M} = \frac{M}{N} = \frac{5,200}{48} = 108.\dot{3}$.

(i)

i	M_i	y_i	$\bar{y}_i = y_i/M_i$
1	161	334	2.0745
2	148	356	2.4054
3	83	245	2.9518
4	157	412	2.6242
5	96	207	2.1563
6	103	315	3.0583
Total	748	1869	15.2705

$$\sum_i M_i = 748;\ \sum_i M_i^2 = 99,188;\ \sum_i \hat{y}_i = 1,869;\ \sum_i \hat{y}_i^2 = 610,135;\ \sum_i \hat{y}_i M_i = 243,798$$

We have $\sum_i \bar{y}_i = 15.2705$, $\sum_i \bar{y}_i^2 = 39.6919$ and $\overline{\overline{y}} = \frac{\sum_i \bar{y}_i}{n} = \frac{15.2705}{6} = 2.545083\dot{3}$.

$$\begin{aligned}
\hat{\overline{\overline{Y}}}_1 &= \overline{\overline{y}} = 2.545083\dot{3} \\
s_{\bar{y}}^2 &= \frac{1}{n-1}(\sum_i \bar{y}_i^2 - n\overline{\overline{y}}^2) = \frac{1}{5}(39.6919 - 6 \times 2.545083\dot{3}^2) = 0.165436365 \\
\mathrm{var}(\hat{\overline{\overline{Y}}}_1) &= \left(1 - \frac{n}{N}\right)\frac{s_{\bar{y}}^2}{n} = \left(1 - \frac{6}{48}\right)\frac{0.165436365}{6} = 0.024126136 \\
\mathrm{se}(\hat{\overline{\overline{Y}}}_1) &= \sqrt{0.024126136} = 0.155325904
\end{aligned}$$

(ii) We have $\bar{y} = \frac{\sum_i y_i}{n} = \frac{1869}{6} = 311.5$. Ordinary estimate for mean per element:

$$\begin{aligned}
\hat{\overline{\overline{Y}}}_2 &= \frac{1}{\overline{\overline{M}}}\bar{y} = \frac{311.5}{108.\dot{3}} = 2.875384815 \\
s_y^2 &= \frac{1}{n-1}(\sum_i y_i^2 - n\bar{y}^2) = \frac{1}{5}(610,135 - 6 \times 311.5^2) = 5,588.3 \\
\mathrm{var}(\hat{\overline{\overline{Y}}}_2) &= \frac{1}{\overline{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = \frac{1}{108.\dot{3}^2}\left(1 - \frac{6}{48}\right)\frac{5,588.3}{6} = 0.069440414 \\
\mathrm{se}(\hat{\overline{\overline{Y}}}_2) &= \sqrt{0.069440414} = 0.263515491
\end{aligned}$$

(iii) Ratio estimate of mean per element:

$$\begin{aligned}
\hat{\overline{\overline{Y}}}_3 &= r = \frac{\sum_i y_i}{\sum_i M_i} = \frac{1,869}{748} = 2.498663102 \\
s_r^2 &= \frac{1}{n-1}(\sum_i y_i^2 - 2r\sum_i M_i y_i + r^2 \sum_i M_i^2) \\
&= \frac{1}{5}(610,135 - 2 \times 2.498663102 \times 243,798 + 2.498663102^2 \times 99,188) \\
&= 2211.804441 \\
\mathrm{var}(\hat{\overline{\overline{Y}}}_3) &= \frac{1}{\overline{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} = \frac{1}{108.\dot{3}^2}\left(1 - \frac{6}{48}\right)\frac{2211.804441}{6} = 0.02748396 \\
\mathrm{se}(\hat{\overline{\overline{Y}}}_3) &= \sqrt{0.02748396} = 0.165782871
\end{aligned}$$

(iv) 1. $\hat{\overline{\overline{Y}}}_3$ is preferred when M is unknown, the variation of M_i is high and M_i is highly and positively correlated to y_i because M is not required in the estimation of $\overline{\overline{Y}}$. Moreover the strong and positive relationship between M_i and y_i is accounted for in the ratio estimate. Moreover $\overline{\overline{Y}}_1$ is a biased estimator similar to the ratio estimator $\overline{\overline{Y}}_3$ but $\overline{\overline{Y}}_2$ is an unbiased estimator.

2. $\overline{\overline{Y}}_2$ is preferred when M is known and the variation of M_i is low because M is required but M_i is not required in the estimation of $\overline{\overline{Y}}$.

4. One-stage cluster sampling is preferred to 2-stage cluster sampling when

   1. The cluster size is small so that sub-sampling is unnecessary and result in too small the sample size.
   2. The variability of elements within cluster is high so that we would like to include all units within cluster into the sample. The usual rule is to have a higher sampling fraction from cluster of higher variability between elements in the cluster.
   3. Easier to implement.

   Two-stage cluster sampling is preferred to 1-stage cluster sampling when

   1. The cluster size is large so that it is infeasible to include all units within cluster into the sample.
   2. If the cluster size varies a lot across clusters, it would be difficult to control the total sample size for a 1-stage cluster sampling.
   3. If the variability of elements within cluster is low, it is unnecessary to include all elements within a cluster into the sample.
   4. If the variability across clusters is high, a 2-stage cluster sample enables a sample of more clusters than a 1-sate cluster sample given the same sample size.