4 Systematic and Cluster Sampling.

4.1 Systematic Sampling

This is a quick and easy method for selecting a sample when the sampling frame is sequenced. Randomly select a start from the first k units where $k = [\frac{N}{n}] + 1$, then select every kth unit thereafter.

The k possible samples will only be of equal size if N is a multiple of k. If $N = 23$ and $n = 5$ then the $k = [\frac{23}{5}] + 1 = 5$ possible samples are units numbered:

S1	S2	S3	S4	S5
y_1	y_2	y_3	y_4	y_5
y_6	y_7	y_8	y_9	y_{10}
y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
y_{16}	y_{17}	y_{18}	y_{19}	y_{20}
y_{21}	y_{22}	y_{23}		

For simplicity we will restrict attention to the case $N = nk$. (If N is large then the following results will be approximately true.) Under systematic sampling only k of the possible $\binom{N}{n}$ samples under SRS are considered. We expect systematic sampling to be beneficial if the k samples are *representative* of the population.

This might be achieved by selecting the sequencing variable carefully. For example, suppose we have a list of businesses giving location, industry, and employment size. We want to estimate the amount of overtime wages paid and we expect such amount to be correlated with the employment size. Then sorting the list according to employment size before drawing the systematic sample allow small, medium and large values to be included in the sample and hence gives more efficient sample.

4.1.1 The Sample Mean and its Variance.

Denote the sample by $y_1, ..., y_n$. Then

$$\bar{y}_{sys} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

can take only one of k possible values:

$$\begin{aligned}
\bar{Y}_1 &= (Y_1 + Y_{k+1} + .. + Y_{N-k+1})/n \\
&\vdots \\
\bar{Y}_k &= (Y_k + Y_{2k} + .. + Y_N)/n.
\end{aligned}$$

Thus

$$E(\bar{y}_{sys}) = (\bar{Y}_1 + ... + \bar{Y}_k)/k = \bar{Y},$$

so the systematic mean is unbiased for the population mean.

Let $Y_{ij} = Y_{i+k(j-1)}$, for $i = 1, 2, ..., k$ and $j = 1, 2, ..., n$.

$$\text{Var}(\bar{y}_{sys}) = E(\bar{y}_{sys} - \bar{Y})^2 = \frac{1}{k}\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})^2.$$

Recall

$$\begin{aligned}
(N-1)S^2 &= \sum_{i=1}^{N}(Y_i - \bar{Y})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2 + n\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})^2 \\
&= k(n-1)S_w^2 + kn\text{Var}(\bar{y}_{sys}) = (N-k)S_w^2 + N\text{Var}(\bar{y}_{sys})
\end{aligned}$$

since $N = nk$. Thus $\text{Var}(\bar{y}_{sys})$ is the *between* cluster variance and S_w^2 is the *within* cluster variance.

$$\text{Var}(\bar{y}_{sys}) = \frac{N-1}{N}S^2 - \frac{N-k}{N}S_w^2$$

Theorem: Systematic sampling leads to more precise estimators of $\bar{Y}$ than SRS if and only if $S_w^2 > S^2$.

Proof: The variance under SRS is $\left(1 - \frac{n}{N}\right)\frac{S^2}{n}$. Since

$$\begin{aligned}
& \text{Var}(\bar{y}) > \text{Var}(\bar{y}_{sys}) \\
\Rightarrow\ & \left(1 - \frac{n}{N}\right)\frac{S^2}{n} > \frac{N-1}{N}S^2 - \frac{N-k}{N}S_w^2 \\
\Rightarrow\ & \frac{N-k}{N}S_w^2 > \left[\frac{N-1}{N} - \frac{N-n}{Nn}\right]S^2 = \frac{1}{nN}(Nn - n - N + n)S^2 \\
& \qquad = \frac{N}{nN}(n-1)S^2 = \frac{N-k}{N}S^2 \\
\Rightarrow\ & S_w^2 > S^2 \qquad (\text{since } N = nk \Rightarrow \frac{1}{n} = \frac{k}{N}).
\end{aligned}$$

Hence Systematic sampling is most useful when the variability within systematic samples is larger than the population variance.

With one systematic sample, we cannot estimate S^2. Instead, we can draw n_s *repeated systematic sample* each of size n/n_s with a random start chosen from 1 to $k' = n_s k = n_s N/n$ and the sampling interval is k'. Suppose $\bar{y}_i, i = 1, ..., n_s$ are the sample means, the mean estimator and its variance are

$$\boxed{\hat{\bar{Y}}_{sy,rs} = \bar{\bar{y}} = \frac{1}{n_s}\sum_{i=1}^{n_s}\bar{y}_i} \quad \text{and} \quad \boxed{\text{var}(\hat{\bar{Y}}_{sy,rs}) = \left(1 - \frac{n}{N}\right)\frac{s_{\bar{y}}^2}{n_s}}$$

where

$$\boxed{s_{\bar{y}}^2 = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(\bar{y}_i - \bar{\bar{y}})^2 = \frac{1}{n_s - 1}\left(\sum_{i=1}^{n_s}\bar{y}_i^2 - n_s\bar{\bar{y}}^2\right).}$$

Read Tutorial 12 Q3.

Systematic Sampling

Parameter	**Point Estimate**	**Estimated Variance**
One systematic sample		
Mean $\overline{Y}$	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\text{Var}(\widehat{\overline{Y}}_{sy}) = \frac{1}{k}\sum_{j=1}^{k}(\bar{y}_j - \bar{y})^2$
(Sampling interval $k = \frac{N}{n}$)		$= \frac{N-1}{N}S^2 - \frac{N-k}{N}S_w^2$
		where $S_w^2 = \frac{1}{N-k}\left[\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \bar{y}_i)^2\right]$
		$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \bar{y})^2\right]$
		$\text{Var}(\widehat{\overline{Y}}_{sy}) < \text{Var}(\widehat{\overline{Y}}_{srs})$ if $S_w^2 > S^2$
Repeated systematic samples		
Mean $\overline{Y}$	$\bar{\bar{y}} = \frac{1}{n_s}\sum_{j=1}^{n_s}\bar{y}_j$	$\text{var}(\widehat{\overline{Y}}_{sy,rs}) = \left(1 - \frac{n}{N}\right)\frac{s_{\bar{y}}^2}{n_s}$
(Sampling interval $k' = \frac{N}{n}\ n_s$)		$s_{\bar{y}}^2 = \frac{1}{n_s - 1}\sum_{j=1}^{n_s}(\bar{y}_j - \bar{\bar{y}})^2$
		$= \frac{1}{n_s - 1}\left(\sum_{j=1}^{n_s}\bar{y}_j^2 - n_s\bar{\bar{y}}^2\right)$

4.2 Cluster Sampling

4.2.1 Introduction

If a population has a natural grouping of units into clusters one way of proceeding is to perform a SRS of the clusters and then include *all* units in the selected clusters in the sample. Systematic sampling has this structure but the distinct 'clusters' may not have any physical significance.

For cluster sampling we do not need a complete population list, just a list of clusters and then a list of elements for the selected clusters.

4.2.2 Notation

N = Number of clusters in the population.
M_i = Size of cluster $i, i = 1, \ldots, N$.
$M = \sum_{i=1}^{N} M_i$ = Population size of elements.
$\bar{M} = \frac{M}{N} = \frac{\sum_{i=1}^{N} M_i}{N}$ = Population average cluster size.
y_{ij} = y-value for element j element in cluster $i, i = 1, .., N; j = 1, .., M_i$.
$Y_i = y_i = \sum_{j=1}^{M_i} y_{ij}$ = y-total for cluster $i, i = 1, \ldots, N$.
$R_i = \bar{\bar{Y}}_i = \frac{y_i}{M_i}$ = y-mean per element for cluster $i, i = 1, \ldots, N$.
$Y = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$ = Population y-total.
$\bar{Y} = \frac{Y}{N}$ = Population mean per cluster.
$R = \bar{\bar{Y}} = \frac{Y}{M}$ = Population mean per element.
n = Number of clusters in the sample.

For 1-stage sampling
$m = \sum_{i=1}^{n} M_i$ = Sample size of elements.
$\bar{m} = \frac{1}{n} \sum_{i=1}^{n} M_i$ = Sample average cluster size.
$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ = Sample mean per cluster.
$r = \bar{\bar{y}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$ = Sample mean per element.

E.g. Sample of clusters in blue

$N = 6,\ n = 3$

$\underline{y_1 = 10},\ \underline{y_2 = 12},\ y_3 = 7,\ y_4 = 9,\ \underline{y_5 = 6},\ y_6 = 7$

$\underline{M = 20},\ m = 11$

$\underline{M_1 = 4},\ \underline{M_2 = 4},\ M_3 = 3,\ M_4 = 2,\ \underline{M_5 = 3},\ M_6 = 4$

$\bar{M} = \frac{20}{6} = 3.33$

$\bar{m} = \frac{11}{3} = 3.67$

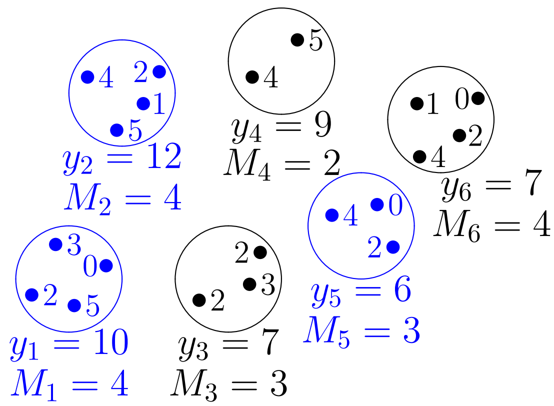
$Y = 51$

$\bar{Y} = \frac{51}{6} = 8.50$

$R = \frac{51}{20} = 2.55$

$\bar{y} = \frac{28}{3} = 9.33$

$r = \frac{28}{11} = 2.55$



	SRS		Cluster sampling
Pop. size	of element	N	$\rightarrow$ of cluster N
Pop. total	of aux. var. X	X	$\rightarrow$ of element M
Sample unit	element with	y_i	$\rightarrow$ cluster with cluster total $y_i = \sum_j y_{ij}$
Auxiliary var.	value	x_i	$\rightarrow$ cluster size M_i
Mean	mean per ele.	$\bar{y}$	$\rightarrow$ mean per cluster $\bar{y}$
Estimator	ordinary	$\bar{y}$	$\rightarrow$ ordinary $\bar{y}$
Estimator	ratio	$r = \frac{\sum_i y_i}{\sum_i x_i}$	$\rightarrow$ ratio $r = \frac{\sum_i y_i}{\sum_i M_i}$

Relationship between estimates

Mean/ele. Mean/cluster Total

$$R \quad \xleftrightarrow{\times \bar{M}} \quad \bar{Y} \quad \xleftrightarrow{\times N} \quad Y$$

$$R \quad \xleftrightarrow{\times M} \quad Y$$

since $\bar{M} \times N = \frac{M}{N} \times N = M$.

4.2.3 Estimators for cluster sampling

Ratio estimate & variance

Mean/ele. R	$\widehat{R}_{c1,r} = r = \dfrac{\bar{y}}{\bar{m}}$		$\text{var}(\widehat{R}_{c1,r}) = \dfrac{1}{\bar{M}^2}(1-\dfrac{n}{N})\dfrac{s_r^2}{n}$
Mean/clus. $\bar{Y}$	$\widehat{\bar{Y}}_{c1,r} = \bar{M}r$	X	$\text{var}(\widehat{\bar{Y}}_{c1,r}) = (1-\dfrac{n}{N})\dfrac{s_r^2}{n}$
Total Y	$\widehat{Y}_{c1,r} = Mr$	X	$\text{var}(\widehat{Y}_{c1,r}) = N^2(1-\dfrac{n}{N})\dfrac{s_r^2}{n}$

Ordinary estimate & variance

Mean/ele. R	$\widehat{R}_{c1} = \dfrac{\bar{y}}{\bar{M}}$	X	$\text{var}(\widehat{R}_{c1}) = \dfrac{1}{\bar{M}^2}(1-\dfrac{n}{N})\dfrac{s_y^2}{n}$
Mean/clus. $\bar{Y}$	$\widehat{\bar{Y}}_{c1} = \bar{y}$		$\text{var}(\widehat{\bar{Y}}_{c1}) = (1-\dfrac{n}{N})\dfrac{s_y^2}{n}$
Total Y	$\widehat{Y}_{c1} = N\bar{y}$		$\text{var}(\widehat{Y}_{c1}) = N^2(1-\dfrac{n}{N})\dfrac{s_y^2}{n}$

Note that X indicates that the estimator cannot be applied when M or $\bar{M}$ is unknown.

Remark:

The ratio estimators use the sample mean per element

$$r = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} = \frac{\sum_{i=1}^n \sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^n M_i}.$$

and the variance estimates follow from the ratio estimators in SRS where

$$s_r^2 = \frac{\sum_{i=1}^n (y_i - rM_i)^2}{n-1} = \frac{\sum_{i=1}^n y_i^2 - 2r\sum_{i=1}^n y_i M_i + r^2 \sum_{i=1}^n M_i^2}{n-1}.$$

Read Tutorial 13 Q1(a).

4.2.4 Equal Cluster Sizes

The analysis is simpler when the clusters are of equal size, that is, $M_i = M/N = \bar{M}, \quad i = 1, 2, .., N$ where M and N are the total number of elements and clusters respectively.

Recall from ANOVA

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(Y_{ij} - \bar{\bar{Y}})^2 &= \sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(Y_{ij} - \bar{\bar{Y}}_i)^2 + \sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2 \\ (M-1)S^2 &= (M-N)S_w^2 + (N-1)S_b^2.\end{aligned}$$

where $\bar{\bar{Y}}_i = Y_i/M_i = R_i$ and $\bar{\bar{Y}} = Y/M = R$ are the *mean per element* for cluster i and overall respectively. The *within* cluster variance is

$$S_w^2 = \frac{1}{M-N}\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(Y_{ij} - \bar{\bar{Y}}_i)^2 = \frac{1}{N(\bar{M}-1)}\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(Y_{ij} - Y_i/M_i)^2.$$

The *between* cluster variance is

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2 = \frac{1}{\bar{M}(N-1)}\sum_{i=1}^{N} M_i\left(\frac{Y_i}{M_i} - \frac{\bar{Y}}{\bar{M}}\right)^2 = \frac{1}{\bar{M}(N-1)}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{S_y^2}{\bar{M}}$$

since $\bar{M} = M_i$ and $R = Y/M = \bar{Y}/\bar{M}$. Based on a SRS of y_i, an unbiased estimator for S_b^2 is

$$s_b^2 = \frac{1}{\bar{M}(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

and an unbiased estimator for S_w^2 is

$$s_w^2 = \frac{1}{n(\bar{M}-1)}\sum_{i=1}^{n}\sum_{j=1}^{\bar{M}}(y_{ij} - \bar{\bar{y}}_i)^2 \text{ where } \bar{\bar{y}}_i = y_i/M_i = r_i.$$

Both components of S^2 can be estimated under cluster sampling unlike systematic sampling where we only observe one 'cluster' and so cannot estimate the between cluster component.

4.2.5 Comparing SRS and Cluster Sampling

A SRS of n clusters includes $n\bar{M}$ observations. An estimate of mean per element $\bar{Y}$ based on one SRS of y_{ij} of the *same size* has variance

$$\text{Var}(\hat{\bar{\bar{Y}}}) = \left(1 - \frac{n\bar{M}}{N\bar{M}}\right)\frac{S^2}{n\bar{M}} = \left(1 - \frac{n}{N}\right)\frac{S^2}{n\bar{M}}.$$

This is compared to a SRS of cluster totals y_i from n clusters

$$\text{Var}(\hat{\bar{\bar{\bar{Y}}}}_{c1}) = \frac{1}{\bar{M}^2}\left(1 - \frac{n}{N}\right)\frac{S_y^2}{n} = \left(1 - \frac{n}{N}\right)\frac{S_b^2}{n\bar{M}}$$

where $\hat{\bar{\bar{\bar{Y}}}}_{c1} = \widehat{R}$ is the mean per element estimate. Thus

$$\frac{\text{Var}(\hat{\bar{\bar{\bar{Y}}}}_{c1})}{\text{Var}(\hat{\bar{\bar{Y}}})} = \frac{S_b^2}{S^2},$$

and so the cluster estimator is preferable when $S_b^2 < S^2$. Note that

$$S_b^2 = \frac{1}{\bar{M}(N-1)}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

is minimised when all the cluster totals are equal, i.e. $Y_i = \bar{Y}, \ i = 1, .., N$, which are then equal to the average cluster total.

In other words, it should be more *homogeneous* across clusters and more *heterogeneous* within each cluster. The worst case occurs when $S_w^2 = 0$ in which case each cluster consists of identical responses.

4.2.6 Two-stage cluster sampling

1. **Description**

   In a single-stage cluster sampling, all the elements in the selected clusters are included in the sample. In practice, it may not be feasible to survey all the elements in the selected clusters when the cluster sizes are large. A natural way is to sub-sample elements from the selected clusters resulting in a 2-stage cluster sampling where at

Stage 1: a SRS of clusters is drawn,

Stage 2: a SRS of elements is taken from each cluster selected at stage 1.

2. **Further notation**

For 2-stage sampling

m_i = Sub-sample size of elements from M_i elements in cluster i.

$m = \sum_{i=1}^{n} m_i$ = Total sample size ($\sum_{i=1}^{n} M_i$ in 1-stage cluster sam.).

$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} M_i$ = Average cluster size in the sample.

$\hat{y}_i = M_i \bar{y}_{i\cdot} = M_i \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i}$ = Estimated y-total for cluster i.

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$ = Sample mean per cluster.

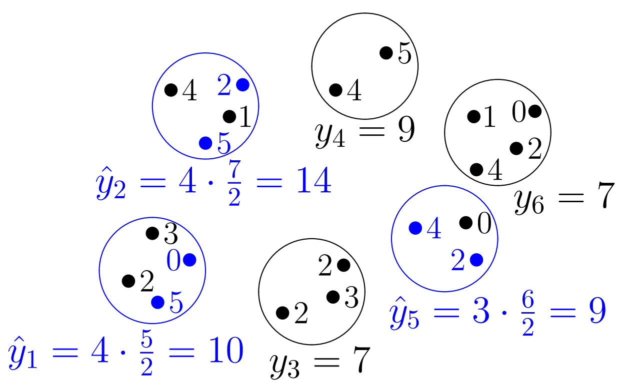
$\hat{r} = \bar{\bar{y}} = \frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} M_i}$ = Sample mean per element.

Comparing 1-stage cluster sample with 2-stage cluster sample:

	1-stage	**2-stage**
Sample size in cluster:	M_i	m_i
Cluster total:	$y_i = \sum_{j=1}^{M_i} y_{ij}$ is observed	$\hat{y}_i = M_i \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i}$ is estimated.
Sample mean per cluster:	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$
Sample mean per element:	$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$	$\hat{r} = \frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} M_i}$

E.g. Sample of clusters and their elements in blue

$N = 6$
$n = 3$
$M = 20$
$m = 6$
$\sum_{i=1}^{n} M_i = 11$
$\bar{M} = \frac{20}{6} = 3.33$
$\bar{m} = \frac{11}{3} = 3.67$
$Y = 51$
$\bar{Y} = \frac{51}{6} = 8.50$
$R = \frac{51}{20} = 2.55$
$\hat{\bar{y}} = \frac{33}{3} = 11$
$\hat{r} = \frac{33}{11} = 3$



3. **Estimation**

The *observed* cluster totals $y_i = \sum_{j=1}^{M_i} y_{ij}$ are estimated by

$$\hat{y}_i = M_i \bar{y}_{i\cdot} = \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} \tag{1}$$

where $\bar{y}_{i\cdot}$ is the sample mean per elements for the elements y_{ij} selected from cluster i at stage 2 and then replace y_i by $\hat{y}_i$.

The additional variance when the cluster total y_i is estimated by $\hat{y}_i = M_i \bar{y}_{i\cdot}$ is

$$\text{var}(\hat{y}_i) = M_i^2\, \text{var}(\bar{y}_{i\cdot}) = M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_{yi}^2}{m_i}$$

where

$$s_{yi}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{i\cdot})^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} y_{ij}^2 - m_i \bar{y}_{i\cdot}^2 \tag{2}$$

is the sample variance for m_i elements y_{ij} selected from cluster i at stage 2. Then for $\widehat{Y} = \sum_{i=1}^{N} \hat{y}_i$ say, estimating the population total $Y = \sum_{i=1}^{N} y_i$, the additional variance, $\sum_{i=1}^{N}$ Add. $\text{var}(\hat{y}_i)$ follows as

$$\sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_{yi}^2}{m_i} \quad : \quad \text{Additional var. due to } \hat{y}_i \text{ for } n \text{ selected clusters}$$

$$\frac{1}{n} \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_{yi}^2}{m_i} \quad : \quad \text{Average additional var. due to } \hat{y}_i \text{ for each cluster}$$

$$\frac{N}{n} \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_{yi}^2}{m_i} \quad : \quad \text{Estimated additional var. due to } \hat{y}_i \text{ for } N \text{ clusters.}$$

Hence the additional variance of $\text{var}(\widehat{\bar{Y}}_{c2,r})$ and $\text{var}(\widehat{\bar{\bar{Y}}}_{c2,r})$ are obtained by dividing the additional variance of $\text{var}(\widehat{Y}_{c2,r})$ with N^2 and M^2 respectively.

For ratio estimator:

$$\text{var}(\widehat{R}_{c2,r}) = \frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{N}{nM^2}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i},$$

$$\text{var}(\widehat{\bar{Y}}_{c2,r}) = \left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i},$$

$$\text{var}(\widehat{Y}_{c2,r}) = N^2\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i}$$

where $\widehat{R} = \hat{r} = \frac{\sum_{i=1}^{n}\hat{y}_i}{\sum_{i=1}^{n}M_i} = \widehat{\bar{\bar{Y}}}_{c2,r}$ is a mean per element estimate and

$$s_r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{r}M_i)^2}{n-1} = \frac{\sum_{i=1}^{n}\hat{y}_i^2 - 2\hat{r}\sum_{i=1}^{n}\hat{y}_iM_i + \hat{r}^2\sum_{i=1}^{n}M_i^2}{n-1} \tag{3}$$

is obtained by substituting $\hat{y}_i$ for y_i in s_r^2 for 1-stage cluster sampling.

For ordinary estimator:

$$\text{var}(\widehat{R}_{c2}) = \frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} + \frac{N}{nM^2}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i},$$

$$\text{var}(\widehat{\bar{Y}}_{c2}) = \left(1-\frac{n}{N}\right)\frac{s_y^2}{n} + \frac{1}{nN}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i},$$

$$\text{var}(\widehat{Y}_{c2}) = N^2\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} + \frac{N}{n}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i}$$

where

$$s_y^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{n-1} = \frac{\sum_{i=1}^{n}\hat{y}_i^2 - n\bar{\hat{y}}^2}{n-1}$$

is obtained by substituting $\hat{y}_i$ for y_i in s_y^2 for 1-stage cluster sampling and $\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$ is the sample mean of $\hat{y}_i$.

Note: It can be shown that $\widehat{R}_{c2}$, $\widehat{\bar{Y}}_{c2}$ and $\widehat{Y}_{c2}$ are exactly unbiased while $\widehat{R}_{c2,r}$, $\widehat{\bar{Y}}_{c2,r}$ and $\widehat{Y}_{c2,r}$ are approximately unbiased for large sample sizes.

Summary

	Ratio **Est.** (S_r^2)	**Ordinary** **Est.** (S_y^2)	**Variance** **in stage 1**	**Additional variance in** **in stage 2** $\hat{y}_i = M_i\bar{\bar{y}}_i \to y_i$
Mean/ele. R	$\hat{r}$ $\surd$	$\frac{\bar{\hat{y}}}{\bar{M}}$ X	$\frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{S^2}{n}$	$\frac{N}{nM^2}\sum_{i=1}^{n}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i}$
Mean/clus. $\bar{Y}$	$\bar{M}\hat{r}$ X	$\bar{\hat{y}}$ $\surd$	$\left(1-\frac{n}{N}\right)\frac{S^2}{n}$	$\frac{1}{nN}\sum_{i=1}^{n}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i}$
Total Y	$M\hat{r}$ X	$N\bar{\hat{y}}$ $\surd$	$N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}$	$\frac{N}{n}\sum_{i=1}^{n}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i}$

Note:

(a) '$\surd$' or 'X' indicates whether we can use the estimator when M is unknown. In the case when M is unknown, $\bar{M}$ is replaced by the sample mean $\bar{m} = \frac{1}{n}\sum_{i=1}^{n}M_i$ and

$$\begin{aligned}
\text{Add. var. for } R &= \frac{N}{nM^2}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i} \\
&= \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i} \\
&\simeq \frac{1}{nN\bar{m}^2}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i}.
\end{aligned}$$

(b) If we replace $\bar{M}$ by $\bar{m}$, $\hat{\bar{\bar{Y}}}_{c2}$ will become $\hat{\bar{\bar{Y}}}_{c2,r}$ since

$$\widehat{R}_{c2} = \frac{\hat{\bar{y}}}{\bar{M}} \to \frac{\hat{\bar{y}}}{\bar{m}} = \frac{\sum_{i=1}^{n}\hat{y}_i}{\sum_{i=1}^{n}M_i} = \hat{r} = \widehat{R}_{c2,r}$$

and $\widehat{Y}_{c2,r}$ will become $\widehat{Y}_{c2}$ since

$$\widehat{Y}_{c2,r} = N\bar{M}\hat{r} \to N\bar{m}\hat{r} = N\bar{m}\frac{\hat{\bar{y}}}{\bar{m}} = N\hat{\bar{y}} = \widehat{Y}_{c2}.$$

Read Tutorial 13 Q1(b).

4.2.7 Stratified one-stage cluster sampling

Stratified cluster sampling can be performed by selecting a cluster sample from each of the L strata in the population. There are ordinary, separate ratio and combined ratio estimators. For the separate ratio estimator, the population size in each stratum N_i are usually unknown, we will investigate only the combined ratio form.

Let y_{li} represents the i-th cluster total in stratum l and $\overline{y}_l$ the average cluster total in stratum l. The following table is the same as the table for stratified sample in P.60.

Stratified one-stage cluster sample

Estimate	Variance
Ordinary estimator: $\bar{y}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} y_{li}$, $s_{ly}^2 = \frac{1}{n_l-1}\sum_{i=1}^{n_l}(y_{li}-\bar{y}_l)^2$	
$\widehat{R}_{stc1} = \frac{1}{\bar{M}}\sum_{l=1}^{L} W_l \bar{y}_l$ X	$\text{var}(\widehat{R}_{stc1}) = \frac{1}{\bar{M}^2}\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{yl}^2}{n_l}$
$\widehat{\bar{Y}}_{stc1} = \sum_{l=1}^{L} W_l \bar{y}_l$ $\surd$	$\text{var}(\widehat{\bar{Y}}_{stc1}) = \sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{yl}^2}{n_l}$
$\widehat{Y}_{stc1} = N\sum_{l=1}^{L} W_l\ \bar{y}_l$ $\surd$	$\text{var}(\widehat{Y}_{stc1}) = N^2\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{yl}^2}{n_l}$
Separate ratio estimator: $r_l = \frac{\sum_{i=1}^{n_l} y_{li}}{\sum_{i=1}^{n_l} M_{li}}$, $s_{srl}^2 = s_{yl}^2 - 2r_l s_{xyl} + r_l^2 s_{xl}^2$	
$\widehat{R}_{stc1,sr} = \frac{1}{\bar{M}}\sum_{l=1}^{L} W_l\ r_l \bar{M}_l$ X	$\text{var}(\widehat{R}_{stc1,sr}) = \frac{1}{\bar{M}^2}\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{srl}^2}{n_l}$
$\widehat{\bar{Y}}_{stc1,sr} = \sum_{l=1}^{L} W_l\ r_l \bar{M}_l$ X	$\text{var}(\widehat{\bar{Y}}_{stc1,sr}) = \sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{srl}^2}{n_l}$
$\widehat{Y}_{stc1,sr} = N\sum_{l=1}^{L} W_l\ r_l \bar{M}_l$ X	$\text{var}(\widehat{Y}_{stc1,sr}) = N^2\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{srl}^2}{n_l}$
Combine ratio estimator: $r_c = \frac{W_1\ \bar{y}_1 + W_2\ \bar{y}_2}{W_1\ \bar{m}_1 + W_2\ \bar{m}_2}$, $s_{crl}^2 = s_{yl}^2 - 2r_c s_{xyl} + r_c^2 s_{xl}^2$	
$\widehat{R}_{stc1,cr} = r_c$ $\surd$	$\text{var}(\widehat{R}_{stc1,cr}) = \frac{1}{\bar{M}^2}\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{crl}^2}{n_l}$
$\widehat{\bar{Y}}_{stc1,cr} = \bar{M}\ r_c$ X	$\text{var}(\widehat{\bar{Y}}_{stc1,cr}) = \sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{crl}^2}{n_l}$
$\widehat{Y}_{stc1,cr} = M\ r_c$ X	$\text{var}(\widehat{Y}_{stc1,cr}) = N^2\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{crl}^2}{n_l}$

'X' indicates that the estimate cannot be applied if $\bar{M}$ or M is unknown.

Example: (City income) Interviews are conducted in each of the 25 blocks sampled from 415 blocks in City 1 with $M_1 = 2500$ individuals and 10 blocks sampled from 168 blocks in City 2 with $M_2 = 850$ individuals. The data on incomes are presented in the table below.

(a) Estimate the overall average income per individual in the two cities and its standard error using

  (i) ordinary estimator if M_1 and M_2 are known,

  (ii) separate ratio estimator if M_1 and M_2 are known,

  (iii) combined ratio estimator if M_1 and M_2 are unknown and

(b) Estimate the average income per household in the two cities combined and its standard error using the four estimators.

(c) Estimate the total income in the two cities combined and its standard error using the four estimators.

City 1										City 2				
i	y_{1i}	M_{1i}	z_{sr1i}	z_{cr1i}	i	y_{1i}	M_{1i}	z_{sr1i}	z_{cr1i}	i	y_{2i}	M_{2i}	z_{sr2i}	z_{cr2i}
1	96000	8	25589	20918	14	49000	10	-39013	-44852	1	18000	2	-4327	-770
2	121000	12	15384	8377	15	53000	9	-26212	-31467	2	52000	5	-3816	5074
3	42000	4	6795	4459	16	50000	3	23596	21844	3	68000	7	-10143	2303
4	65000	5	20993	18074	17	32000	6	-20808	-24311	4	36000	4	-8653	-1541
5	52000	6	-808	-4311	18	22000	5	-22007	-24926	5	45000	3	11510	16844
6	40000	6	-12808	-16311	19	45000	5	993	-1926	6	96000	8	6694	20918
7	75000	7	13391	9303	20	37000	4	1795	-541	7	64000	6	-2980	7689
8	65000	5	20993	18074	21	51000	6	-1808	-5311	8	115000	10	3367	21148
9	45000	8	-25411	-30082	22	30000	8	-40411	-45082	9	41000	3	7510	12844
10	50000	3	23596	21844	23	39000	7	-22609	-26697	10	12000	1	837	2615
11	85000	2	67397	66230	24	47000	3	20596	18844					
12	43000	6	-9808	-13311	25	41000	8	-29411	-34082					
13	54000	5	9993	7074										
M						53160	6.04	0.00	-3526.90		54700	4.90	0.00	8712
s						21784	2.37	25189	25999		32342	2.85	7193	8656

Note: $z_{srli} = y_{li} - r_l M_{li}$, $z_{crli} = y_{li} - r_c M_{li}$, $\hat{\rho}_1 = 0.30315$ and $\hat{\rho}_2 = 0.97498$.

Solution: Note that

	n	N	m	M	Mean	SD
Income for city 1, y_{1i}	25	415	151	2500	$\overline{y}_1 = 53,160$	$s_{y1} = 21,784.322$
Resident for city 1, M_{1i}					$\overline{m}_1 = 6.04$	$s_{M1} = 2.3714$
Col. $z_{sr1i} = y_{1i} - r_1 M_{1i}$					$\overline{z}_{sr1} = 0$	$s_{sr1} = 25,189.308$
Col. $z_{cr1i} = y_{1i} - r_c M_{1i}$					$\overline{z}_{cr1} = -3,526.90$	$s_{cr1} = 25,998.656$
Income for city 2, y_{2i}	10	168	49	850	$\overline{y}_2 = 54,700$	$s_{y2} = 32,342.10$
Resident for city 2, M_{2i}					$\overline{m}_2 = 4.90$	$s_{M2} = 2.846$
Col. $z_{sr2i} = y_{2i} - r_2 M_{2i}$					$\overline{z}_{sr2} = 0$	$s_{sr2} = 7,192.970$
Col. $z_{cr2i} = y_{2i} - r_c M_{2i}$					$\overline{z}_{cr2} = 8,712.28$	$s_{cr2} = 8,656.483$
Total	35	583	200	3350		

(a) We have

$$r_1 = \frac{\overline{y}_1}{\overline{m}_1} = \frac{53,160}{6.04} = 8,801.325 \quad r_2 = \frac{\overline{y}_2}{\overline{m}_2} = \frac{54,700}{4.90} = 11,163.265 \quad \textit{sep. r., mean/ele.}$$

$$r_c = \frac{N_1\overline{y}_1 + N_2\overline{y}_2}{N_1\overline{m}_1 + N_2\overline{m}_2} = \frac{415 \times 53,160 + 168 \times 54,700}{415 \times 6.04 + 168 \times 4.90} = 9,385.248 \quad \textit{comb. r., mean/ele.}$$

$$\overline{M}_1 = \frac{M_1}{N_1} = \frac{2,500}{415} = 6.024 \qquad \overline{M}_2 = \frac{M_2}{N_2} = \frac{850}{168} = 5.060 \qquad \textit{pop. mean cluster size}$$

$$\overline{M} = \frac{M}{N} = \frac{3,350}{583} = 5.7461 \qquad \textit{overall pop. mean cluster size}$$

$$W_1 = \frac{N_1}{N} = \frac{415}{583} = 0.7118 \qquad W_2 = \frac{N_2}{N} = \frac{168}{583} = 0.2882 \qquad \textit{wt. prop. to no. of cluster}$$

$$
\begin{aligned}
s_{sr1}^2 &= s_{y1}^2 - 2r_1\rho_1 s_{y1}s_{M1} + r_1^2 s_{M1}^2 \\
&= 21784.322^2 - 2(8801.325)(0.30315)(21784.322)(2.3714) + (8801.325^2(2.3714^2) \\
&= 25189.308^2 \\
s_{cr1}^2 &= s_{y1}^2 - 2r_c\rho_1 s_{y1}s_{M1} + r_c^2 s_{M1}^2 \\
&= 21784.322^2 - 2(9385.248)(0.30315)(21784.322)(2.3714) + 9385.248^2(2.3714^2) \\
&= 25998.656^2 \\
s_{sr2}^2 &= s_{y2}^2 - 2r_2\rho_2 s_{y2}s_{M2} + r_2^2 s_{M2}^2 \\
&= 32342.1^2 - 2(11163.265)(0.97498)(32342.1)(2.846) + 11163.265^2(2.846^2) \\
&= 7192.97^2 \\
s_{cr2}^2 &= s_{y2}^2 - 2r_c\rho_2 s_{y2}s_{M2} + r_c^2 s_{M2}^2 \\
&= 32342.1^2 - 2(9385.248)(0.97498)(32342.1)(2.846) + 9385.248^2(2.846^2) \\
&= 8656.483^2
\end{aligned}
$$

(i) Estimate of the average income per individual in the two cities combined and its standard error using *ordinary* estimator are

$$
\begin{aligned}
\widehat{R}_{stc1} &= \frac{\text{sample average of income per block}}{\text{population average no. of residents per block}} \\
&= \frac{1}{\bar{M}}(W_1\overline{y}_1 + W_2\overline{y}_2) \\
&= \\
&= \\
\text{var}(\widehat{R}_{stc1}) &= \frac{1}{\bar{M}^2}\left[W_1^2\left(1-\frac{n_1}{N_1}\right)\frac{s_{y1}^2}{n_1} + W_2^2\left(1-\frac{n_2}{N_2}\right)\frac{s_{y2}^2}{n_2}\right] \\
&= \\
&= \\
se(\widehat{R}_{stc1}) &=
\end{aligned}
$$

(ii) Estimate of the average income per individual in the two cities combined and its standard error using *separate ratio* estimator are

$$\begin{aligned}
\widehat{R}_{stc1,sr} &= \frac{\text{sample average of income per block}}{\text{population average no. of residents per block}} \\
&= \frac{1}{\bar{M}}(W_1\bar{M}_1 r_1 + W_2\bar{M}_2 r_2) \\
&= \\
&=
\end{aligned}$$

$$\begin{aligned}
\text{var}(\widehat{\overline{Y}}_{stc1,sr}) &= \frac{1}{\bar{M}^2}\left[W_1^2\left(1-\frac{n_1}{N_1}\right)\frac{s_{r1}^2}{n_1} + W_2^2\left(1-\frac{n_2}{N_2}\right)\frac{s_{r2}^2}{n_2}\right] \\
&= \\
&= \\
&= \\
se(\widehat{\overline{Y}}_{stc1,sr}) &=
\end{aligned}$$

(iii) Since M is unknown, it is estimated by

$$\widehat{M} = N_1\overline{m}_1 + N_2\overline{m}_2 =$$

Estimate of the average income per individual in the two cities combined and its standard error using *combined ratio* estimator are

$$
\begin{aligned}
\widehat{R}_{stc1,cr} &= \frac{\text{sample estimate of total income}}{\text{sample estimate of total no. of residents}} \\
&= r_c = \frac{N_1\overline{y}_1 + N_2\overline{y}_2}{N_1\overline{m}_1 + N_2\overline{m}_2} \\
&= \\
\text{var}(\widehat{R}_{stc1,cr}) &= \frac{1}{\widehat{M}^2}\left[N_1^2\left(1-\frac{n_1}{N_1}\right)\frac{s_{cr1}^2}{n_1} + N_2^2\left(1-\frac{n_2}{N_2}\right)\frac{s_{cr2}^2}{n_2}\right] \\
&= \\
&= \\
se(\widehat{R}_{stc1,cr}) &=
\end{aligned}
$$

Note that if $M = 3350$ is known, it should be used for $\widehat{M} = 3329.8$ instead and $\text{var}(\widehat{R}_{stc1,cr}) = 402165.9266$ & $\text{se}(\widehat{R}_{stc1,cr}) = 634.1655$ respectively.

(b) , (c) Left as exercise.

Note that

1. The 3 estimates, $\widehat{R}_{stc1} = 9,328.66$, $\widehat{R}_{stc1,sr} = 9,400.62$ and $\widehat{R}_{stc1,cr} = 9,385.25$ using respectively ordinary, separate ratio and combined ratio estimators are similar.

2. Since

$$0.30315 = \rho_1 < \frac{\bar{y}_1 s_{1m}}{2\bar{m}_1 s_{1y}} = \frac{53160(2.3714)}{2(6.04)(21784.322)} = 0.479048$$

$s_{1sr} > s_{1y}$. Also since $\text{se}(\widehat{\overline{Y}}_{stc1,sr}) = 605.755 < \text{se}(\widehat{R}_{stc1,cr}) = 638.013 < \text{se}(\widehat{R}_{stc1}) = 721.92$, the ratio estimate is better than the ordinary estimate because the block sizes M_{2j} are highly and positively related to the total incomes y_{2j} for block j of city 2.

3. Between the two ratio estimates, the separate ratio estimate $\widehat{R}_{stc1}$ is again better because there is great difference between the two ratios of average income per residents $r_1 = 8,801.3245$ and $r_2 = 11,163.265$ in the 2 cities and their sample sizes of blocks, $n_1 = 25$ and $n_2 = 10$ are not too small. Hence the assumption of a common ratio in the combined ratio estimator is not appropriate.