# CHAPTER 12

# Cluster Sampling

## 12.1 INTRODUCTION

Suppose we want to select a sample of 2000 matric students from all the matric students in South Africa to estimate the average cost of education. In this case if we use the simple random sampling with replacement (SRSWR), simple random sampling without replacement (SRSWOR), or systematic sampling procedures for selecting the sample, we will first require a sampling frame that comprises a list of all the matric students in South Africa. In most situations such a list is not readily available. We can then look for a list of all the schools in South Africa that offer a matric program. Such a list is much easier to obtain and we may select a sample of schools, instead of selecting students directly, and then survey all the matric students of the selected schools. This type of sampling procedure is called cluster sampling. Technically, a group of units is called a cluster and each unit in a cluster is called an ultimate unit (from which statistical information is to be collected). In cluster sampling, entire population is first divided into a number of mutually exclusive and exhaustive clusters. We then select a sample of clusters by some suitable method of sampling such as SRSWR, SRSWOR, or probability proportional to size with replacement sampling (PPSWR) and survey all the ultimate units belonging to the sampled clusters.

Cluster sampling may be used when the sampling frame of the ultimate units is not readily available or expensive to construct. This method of sampling is much cheaper, easier, and operationally convenient because units within the same cluster are generally located close to each other and hence traveling costs are much less than that of a sampling procedure, which selects units directly from the entire population.

As per efficiency (sampling variance) is concerned, cluster sampling is generally less efficient than a sampling scheme, which selects units directly. Furthermore, when the cluster sizes are unknown or unequal, we cannot select a desired number of ultimate units.

*Survey Sampling Theory and Applications*
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00012-1

## 12.2 ESTIMATION OF POPULATION TOTAL AND VARIANCE

Consider a finite population that is divided into $N$ mutually exclusive and exhaustive clusters. Let $M_i$ be the size of the $i$th cluster, the number of ultimate units that belong to the $i$th cluster, $i = 1,\ldots, N$. The total number of ultimate units in the population is $M = \sum_{i=1}^{N} M_i$. Let $y_{ij}$ be the value of the character under study for the $j$th ultimate unit of the $i$th cluster $j = 1,\ldots, M_i$; $i = 1,\ldots, N$, $Y_i = \sum_{j=1}^{M_i} y_{ij} = i$th cluster total, and $Y = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$ = population total. Suppose a sample $s$ of size $n$ clusters is selected with probability $p(s)$ by using some suitable sampling design and all the ultimate units that belong to the selected clusters are surveyed. For example, if the $k$th cluster is selected in the sample $s$, i.e., if $k \in s$, then values $y_{kj}$, $j = 1,\ldots, M_k$ and hence the total $Y_k = \sum_{j=1}^{M_k} y_{kj}$ is obtained from the survey.

An unbiased estimator of the population total $Y$ is given by

$$\widehat{Y}_{cs} = \sum_{i \in s} b_{si} Y_i \tag{12.2.1}$$

where $b_{si}$'s are known constants that satisfy the unbiasedness condition.

$$\sum_{s \supset i} b_{si} p(s) = 1 \tag{12.2.2}$$

Theorem 12.2.1

(i) The variance of $\widehat{Y}_{cs}$ is $V\left(\widehat{Y}_{cs}\right) = \sum_{i=1}^{N} \alpha_i Y_i^2 + \sum_{i \neq}^{N}\sum_{j=1}^{N} \alpha_{ij} Y_i Y_j$

where $\alpha_i = \sum_{s \supset i} b_{si}^2 p(s) - 1$, $\alpha_{ij} = \sum_{s \supset i} b_{si} b_{sj} p(s) - 1$

(ii) An unbiased estimator of the variance of $\widehat{Y}_c$ is

$$\widehat{V}\left(\widehat{Y}_c\right) = \sum_{i \in s} c_{si} Y_i^2 + \sum_{i \neq} \sum_{j \in s} c_{sij} Y_i Y_j$$

where $c_{si}$ and $c_{sij}$ are suitably chosen constants that satisfy $\sum_{s \supset i,j} c_{si} p(s) = \alpha_i$ and $\sum_{s \supset i,j} c_{sij} p(s) = \alpha_{ij}$.

Proof

(i) $V\left(\widehat{Y}_{cs}\right) = E\left(\widehat{Y}_{cs}\right)^2 - Y^2$

$$= \sum_s \left( \sum_{i \in s} b_{si}^2 Y_i^2 + \sum_{i \neq} \sum_{j \in s} b_{sij} Y_i Y_j \right) p(s) - Y^2$$

$$= \sum_{i=1}^{N} Y_i^2 \left( \sum_{s \supset i} b_{si}^2 p(s) - 1 \right) + \sum_{i \neq}^{N} \sum_{j=1}^{N} Y_i Y_j \sum_{s \supset} \sum_{i,j} (b_{sij} p(s) - 1)$$

$$= \sum_{i=1}^{N} \alpha_i Y_i^2 + \sum_{i \neq}^{N} \sum_{j=1}^{N} \alpha_{ij} Y_i Y_j$$

(ii) $E\left[\widehat{V}\left(\widehat{Y}_{cs}\right)\right] = E\left( \sum_{i \in s} c_{si} Y_i^2 \right) + E\left( \sum_{i \neq} \sum_{j \in s} c_{sij} Y_i Y_j \right)$

$$= \sum_{i=1}^{N} Y_i^2 \left( \sum_{s \supset i} c_{si} p(s) \right) + \sum_{i \neq}^{N} \sum_{j=1}^{N} Y_i Y_j \left( \sum_{s \supset} \sum_{i,j} c_{sij} p(s) \right)$$

$$= V\left(\widehat{Y}_{cs}\right)$$

Remark 12.2.1

We can choose $c_{si}$ and $c_{sij}$ in various ways. The obvious choices are $c_{si} = 1/\pi_i$ and $c_{sij} = 1/\pi_{ij}$, where $\pi_i$ and $\pi_{ij}(>0)$ are the inclusion probabilities of the $i$th and $i$th and $j$th $(i \neq j)$ clusters, which are assumed to be positive.

## 12.2.1 Arbitrary Sampling Design

Let a sample $s$ be selected by using a fixed effective size sampling design of size $n$ and let $\pi_i$ and $\pi_{ij}$ be the inclusion probabilities of $i$th and $i$th and $j$th units, respectively. Then the following results are obtained from Theorem 12.2.1.

Theorem 12.2.2

(i) $\widehat{Y}_{cs}(ht) = \sum_{i \in s} \frac{Y_i}{\pi_i}$ is an unbiased estimator for the population total $Y$

(ii) $V\left[\widehat{Y}_{cs}(ht)\right] = \frac{1}{2} \sum_{i \neq} \sum_{j \in U} \Delta_{ij} \left( \frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$ and

(iii) $\widehat{V}\left[\widehat{Y}_{cs}(ht)\right] = \frac{1}{2} \sum_{i \neq} \sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \left( \frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$

where $\Delta_{ij} = \pi_i \pi_j - \pi_{ij}$.

### 12.2.2 Simple Random Sampling Without Replacement

For an SRSWOR sampling, $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/\{N(N-1)\}$. Substituting $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/\{N(N-1)\}$ in Theorem 12.2.2, we obtain

Theorem 12.2.3

(i) $\widehat{Y}_{cs}(wor) = N\widehat{\overline{Y}}_{cs}$ is an unbiased estimator for the population total $Y$

(ii) $V\left[\widehat{Y}_{cs}(wor)\right] = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S_Y^2$ and

(iii) $\widehat{V}\left[\widehat{Y}_{cs}(wor)\right] = N^2\left(\frac{1}{n} - \frac{1}{N}\right)s_Y^2$

where

$$\widehat{\overline{Y}}_{cs} = \sum_{i\in s} Y_i/n,\ S_Y^2 = \sum_{i=1}^{N}\left(Y_i - \overline{Y}_{cs}\right)^2/(N-1),\ \overline{Y}_{cs} = Y/N,$$

$$s_Y^2 = \sum_{i\in s}\left(Y_i - \widehat{\overline{Y}}_{cs}\right)^2/(n-1).$$

Now writing $\overline{M} = \sum_{i=1}^{N} M_i/N =$ average cluster size and $\overline{\overline{Y}} = Y/\left(N\overline{M}\right) = \overline{Y}_{cs}/\overline{M} =$ mean per unit, we get

$$\begin{aligned}(N-1)S_Y^2 &= \sum_{i=1}^{N}\left(Y_i - \overline{M}\,\overline{\overline{Y}}\right)^2 \\ &= \sum_{i=1}^{N}\left(\sum_{j=1}^{M_i}\left(y_{ij} - \overline{\overline{Y}}\right) + \left(M_i - \overline{M}\right)\overline{\overline{Y}}\right)^2 \\ &= \sum_{i=1}^{N}\left[\sum_{j=1}^{M_i}\left(y_{ij} - \overline{\overline{Y}}\right)^2 + \sum_{j\neq}^{M_i}\sum_{k=1}^{M_i}\left(y_{ij} - \overline{\overline{Y}}\right)\left(y_{ik} - \overline{\overline{Y}}\right)\right. \\ &\qquad \left. + \overline{\overline{Y}}^2\left(M_i - \overline{M}\right)^2 + 2\overline{\overline{Y}}\ M_i\left(M_i - \overline{M}\right)\left(\overline{Y}_i - \overline{\overline{Y}}\right)\right] \\ &= \left(N\overline{M} - 1\right)S_y^2\left[1 + \rho_c\frac{\sum_{i=1}^{N} M_i(M_i-1)}{N\overline{M}}\right] + Q\end{aligned} \tag{12.2.3}$$

where

$$S_y^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(y_{ij} - \overline{\overline{Y}}\right)^2 \Big/ \left(N\overline{M} - 1\right),$$

$$Q = \overline{\overline{Y}}^2 \sum_{i=1}^{N}\left(M_i - \overline{M}\right)^2 + 2\overline{\overline{Y}}\sum_{i\in U} M_i\left(M_i - \overline{M}\right)\left(\overline{Y}_i - \overline{\overline{Y}}\right)$$

and

$$\rho_c = \frac{\sum_{i=1}^{N}\sum_{j\neq}^{M_i}\sum_{k=1}^{M_i}\left(y_{ij} - \overline{\overline{Y}}\right)\left(y_{ik} - \overline{\overline{Y}}\right) \Big/ \sum_{i=1}^{N} M_i(M_i - 1)}{\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(y_{ij} - \overline{\overline{Y}}\right)^2 \Big/ \sum_{i=1}^{N} M_i}$$

= intracluster correlation coefficient between the elements within clusters.

Now substituting Eq. (12.2.3) in the expression of $V\left(\widehat{Y}_{cs}\right)$ of the Theorem 12.2.3, an alternative expression of $V\left(\widehat{Y}_{cs}\right)$ comes out as

$$V\left(\widehat{Y}_{cs}(wor)\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1} \times \left[\left(N\overline{M} - 1\right)S_y^2\left(1 + \rho_c\frac{\sum_{i=1}^{N} M_i(M_i - 1)}{N\overline{M}}\right) + Q\right] \tag{12.2.4}$$

In the case that clusters are of equal size, i.e., $M_i = \overline{M}$, $Q$ becomes equal to zero and Eq. (12.2.4) reduces to

$$V\left(\widehat{Y}_{cs}(wor)\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{\left(N\overline{M} - 1\right)}{N-1}S_y^2\left[1 + \left(\overline{M} - 1\right)\rho_c\right] \tag{12.2.5}$$

## 12.3 EFFICIENCY OF CLUSTER SAMPLING

Instead of using cluster sampling, if one selects an SRSWOR sample of size $n\overline{M}$ from a population of size $N\overline{M}$, then the variance of the population total $\widehat{Y} = N\ \overline{y}_s$ ($\overline{y}_s$ = sample mean) is obtained as

$$V\left(\widehat{Y}\right) = \left(N\overline{M}\right)^2\left(\frac{1}{n\overline{M}} - \frac{1}{N\overline{M}}\right)S_y^2 \tag{12.3.1}$$

where $(NM - 1)S_y^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(y_{ij} - \overline{\overline{Y}}\right)^2$.

The efficiency of cluster sampling compared with SRSWOR sampling based on the same sample size $n\overline{M}$ is given by

$$E_{cs} = \frac{V\left(\widehat{Y}\right)}{V\left[\widehat{Y}_{cs}(wor)\right]} \tag{12.3.2}$$

In case $M_i = \overline{M}$ for $i = 1,\ldots, N$, and $N$ is very large, the efficiency $E_{cs}$ reduces to

$$E_{cs} = \frac{1}{1 + \left(\overline{M} - 1\right)\rho_c} \tag{12.3.3}$$

From the expression (Eq. 12.3.3), we conclude that the cluster sampling under SRSWOR will be more, equally, or less efficient than the unistage sampling (direct sampling) based on an SRSWOR of the same sample size if the intracluster correlation coefficient $\rho_c$ is negative, zero, or positive, respectively. Clusters are generally formed by taking neighboring units that are expected to produce positive intraclass correlation; hence cluster sampling is generally less efficient than SRSWOR sampling of the same sample size. But as per cost of the survey is concerned, cluster sampling is much cheaper for surveying neighboring units. However, if $\rho_c$ is negative, cluster sampling is favorable in terms of both cost and efficiency.

Because $\rho_c$ is generally positive, the efficiency of cluster sampling $E_{cs}$ decreases with the increase in the cluster size $\overline{M}$. Because the number of clusters decreases with the increase in the size of the clusters, the efficiency of cluster sampling increases with the number of clusters. The cost of the survey, however, rises with the increase in the total number of clusters.

### 12.3.1 Optimum Choice of Cluster Size

As stated previously, if the total sample size is fixed, the efficiency of the cluster sampling increases with the number of clusters, but the cost of the survey also increases with the number of clusters. Because efficiency and cost are moving in opposite directions, a balance between the cost of the survey and efficiency should be established by choosing the cluster size $\overline{M}$ optimally. In determination of the optimum cluster size, let us assume $M_i = M = \overline{M}$ for $i = 1,\ldots, N$ and the cost of the survey is of the form

$$C = c_0 + c_1 Mn + c_2\sqrt{n} \tag{12.3.4}$$

where $c_0$ is the fixed overhead cost, $c_1$ is the cost of surveying a unit, and $c_2\sqrt{n}$ is the total cost of traveling from one cluster to the other, which is

assumed to be approximately proportional to the square root of the number of clusters.

Here we determine the optimum values of $M$ and $n$, which minimize the variance of $\hat{\overline{\overline{Y}}}_{cs}(wor) = \dfrac{\hat{Y}_{cs}(wor)}{MN}$, the estimator of the mean per unit keeping the total cost of the survey $C$ fixed to a certain level $C_0$. For this minimization problem consider

$$\phi = V\left[\hat{\overline{\overline{Y}}}_{cs}(wor)\right] + \lambda\left(C^* - c_1 Mn - c_2\sqrt{n}\right) \tag{12.3.5}$$

where $C^* = C_0 - c_0$ and $\lambda$ is a Lagrange multiplier.

Now

$$\begin{aligned} V\left[\hat{\overline{\overline{Y}}}_{cs}(wor)\right] &= \frac{1}{M^2}\left(\frac{1}{n} - \frac{1}{N}\right)\ \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{\overline{Y}}\right)^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right)\ \frac{1}{N-1}\sum_{i=1}^{N}\left(\overline{Y}_i - \overline{\overline{Y}}\right)^2 \end{aligned} \tag{12.3.6}$$

(noting $M_i = M$ for $i = 1,\ldots, N$).

Because the expression (Eq. 12.3.6) is not an explicit function of $M$, the minimization of $\phi$ in Eq. (12.3.5) with respect to $M$ is not possible. However, Smith (1938), Mahalanobis (1940, 1942), Jessen (1942), Hendricks (1944), and Sukhatme et al. (1984), among others, modeled within-cluster mean square $S_w^2 = \dfrac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \overline{Y}_i\right)^2}{N(M-1)}$ as an explicit function of $M$. Empirical investigations based on agricultural surveys reveal that $S_w^2$ can be modeled as

$$S_w^2 = \alpha\ M^g$$

where $\alpha(>0)$ and $g(\geq 0)$ are independent of $M$.

Now writing $M\sum_{i=1}^{N}\left(\overline{Y}_i - \overline{\overline{Y}}\right)^2 = (NM-1)S_y^2 - N(M-1)S_w^2 = (NM-1)S_y^2 - N(M-1)\alpha M^g$ and neglecting finite population correction term we get

$$\phi = \frac{1}{n}\left[S_y^2 - (M-1)\alpha M^{g-1}\right] + \lambda\left(C_0 - c_0 - c_1 Mn - c_2\sqrt{n}\right) \tag{12.3.7}$$

To find the optimum values of $m$ and $n$, we first find the optimum value of $n$ from the equation $C^* = c_1 Mn + c_2\sqrt{n}$ keeping $M$ fixed. The optimum value of $n$ is obtained as

$$n = \left(\frac{-c_2 + \sqrt{c_2^2 + 4c_1 C^* M}}{2c_1 M}\right)^2 \tag{12.3.8}$$

Differentiating $\phi$ with respect to $m$ and $n$ and then equating them to zero we get

$$\frac{\partial \phi}{\partial M} = -\frac{\alpha M^{g-2}}{n}[gM - (g-1)] - \lambda c_1 n = 0 \tag{12.3.9}$$

and

$$\frac{\partial \phi}{\partial n} = -\frac{1}{n^2}\left[S_y^2 - (M-1)\alpha M^{g-1}\right] + \lambda\left(-c_1 M - \frac{c_2}{2\sqrt{n}}\right) = 0 \tag{12.3.10}$$

Eliminating $\lambda$ from Eqs. (12.3.9) and (12.3.10), we have

$$\frac{\alpha M^{g-1}\{gM - (g-1)\}}{S_y^2 - (M-1)\alpha M^{g-1}} = \left(1 + \frac{c_2}{2c_1 M\sqrt{n}}\right)^{-1} \tag{12.3.11}$$

Now substituting $\sqrt{n}$ from Eq. (12.3.8) in Eq. (12.3.11), we get

$$\frac{\alpha M^{g-1}\{gM - (g-1)\}}{S_y^2 - (M-1)\alpha M^{g-1}} = 1 - \frac{c_2}{\sqrt{c_2^2 + 4c_1 C^* M}} \tag{12.3.12}$$

Because Eq. (12.3.12) is independent of $n$, we can obtain $M$ numerically by an iterative procedure. The optimum value of $n$ is obtained from Eq. (12.3.8) by substituting the optimum value of $M$ obtained from Eq. (12.3.12).

## 12.4 PROBABILITY PROPORTIONAL TO SIZE WITH REPLACEMENT SAMPLING

Let a sample $s$ of $n$ clusters be selected by PPSWR method of sampling using normed size measure $p_i\left(>0, \sum_{i=1}^{N} p_i = 1\right)$ attached to the $i$th unit. The Hansen and Hurwitz (1943) estimator for the population total $Y$ is

$$\widehat{Y}_{cs}(hh) = \frac{1}{n}\sum_{r=1}^{n}\frac{Y(r)}{p(r)} \tag{12.4.1}$$

where $Y(r)$ is the cluster total of the unit selected at the $r$th draw with probability $p(r)$, i.e., if the $r$th draw produces the $j$th cluster then $Y(r) = Y_j$ and $p(r) = p_j$.

Theorem 12.4.1

(i) $E\left[\widehat{Y}_{cs}(hh)\right] = Y$

(ii) $V\left[\widehat{Y}_{cs}(hh)\right] = \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{Y_i}{p_i} - Y\right)^2$

(iii) $\widehat{V}\left[\widehat{Y}_{cs}(hh)\right] = \frac{1}{n(n-1)}\sum_{r=1}^{n}\left(\frac{Y(r)}{p(r)} - \widehat{Y}_{cs}(hh)\right)^2$

Proof

This theorem follows directly from Theorem 5.2.2 and hence it is omitted.

### 12.4.1 Simple Random Sampling With Replacement

Suppose a sample of size $n$ is selected by SRSWR method, then an unbiased estimator of the population total $Y$ is given by

$$\widehat{Y}_{cs}(wr) = \frac{N}{n}\sum_{r=1}^{n} Y(r) \tag{12.4.2}$$

where $Y(r)$ is the cluster selected at the $r$th draw. We then have the following theorem by substituting $p_i = 1/N$ for $i = 1,\ldots, N$ in Theorem 12.4.1.

Theorem 12.4.2

(i) $E\left[\widehat{Y}_{cs}(wr)\right] = Y$

(ii) $V\left[\widehat{Y}_{cs}(wr)\right] = N^2\sigma_Y^2/n$

(iii) $\widehat{V}\left[\widehat{Y}_{cs}(wr)\right] = N^2\widehat{\sigma}_Y^2/n$

where $\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$, $\widehat{\sigma}_Y^2 = \frac{1}{(n-1)}\sum_{r=1}^{n}\left[Y(r) - \widehat{Y}_{cs}(wr)\right]^2$, and $\overline{Y} = Y/N$.

## 12.5 ESTIMATION OF MEAN PER UNIT

The mean per unit is defined by $\overline{\overline{Y}} = Y/\left(N\overline{M}\right)$. Hence if $\overline{M}$ is known, $\overline{\overline{Y}}$ may be estimated unbiasedly by

$$\widehat{\overline{\overline{Y}}}_{cs} = \widehat{Y}_{c}/\left(N\overline{M}\right) \tag{12.5.1}$$

In practice $\overline{M}$ is unknown. Hence $\overline{\overline{Y}}$ may be estimated through a ratio estimator as follows:

$$\widehat{\overline{\overline{Y}}}_{Rc} = \widehat{Y}_{cs}/\widehat{M}_{cs} \tag{12.5.2}$$

where $\widehat{Y}_{cs} = \sum_{i\in s} b_{si}Y_i$ and $\widehat{M}_{cs} = \sum_{i\in s} b_{si}M_i$.

The ratio estimator $\hat{\bar{\bar{Y}}}_{Rc}$ is biased for $\bar{\bar{Y}}$. However, the amount of bias is negligible for a large sample. Approximate expressions of the bias and the mean square error of $\hat{\bar{\bar{Y}}}_{Rc}$ are obtained from Eqs. (8.2.11) and (8.2.16), respectively, as follows:

$$B\left(\hat{\bar{\bar{Y}}}_{Rc}\right) = -\frac{Cov\left(\hat{M}_{cs}, \hat{G}_{cs}\right)}{\left(N\bar{M}\right)^2} \tag{12.5.3}$$

[where $\hat{G}_{cs} = \sum_{i \in s} b_{si} G_i$ and $G_i = M_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)$] and

$$\begin{aligned} M\left(\hat{\bar{\bar{Y}}}_{Rc}\right) &= \frac{1}{\left(N\bar{M}\right)^2} V\left(\sum_{i \in s} b_{si} G_i\right) \\ &= \frac{1}{\left(N\bar{M}\right)^2}\left(\sum_{i \in U} \alpha_i G_i^2 + \sum_{i \neq} \sum_{j \in U} \alpha_{ij} G_i G_j\right) \end{aligned} \tag{12.5.4}$$

where $\alpha_i = \sum_{s \supset i} b_{si}^2 p(s) - 1$, $\alpha_{ij} = \sum_{s \supset i,j} b_{si} b_{sj} p(s) - 1$, and $p(s)$ is the probability of selection of the sample $s$.

Using Eqs. (8.2.13) and (8.2.19), approximate unbiased estimators of $B\left(\hat{\bar{\bar{Y}}}_{Rc}\right)$ and $M\left(\hat{\bar{\bar{Y}}}_{Rc}\right)$ are obtained, respectively, as

$$\hat{B}\left(\hat{\bar{\bar{Y}}}_{Rc}\right) = -\frac{\hat{C}ov\left(\hat{M}_{cs}, \hat{G}_{cs}\right)}{\left(\hat{M}_{cs}\right)^2}$$

and

$$\hat{M}\left(\hat{\bar{\bar{Y}}}_{Rc}\right) = \frac{1}{\left(\hat{M}_{cs}\right)^2}\left(\sum_{i \in s} c_{si} \hat{G}_i^2 + \sum_{i \neq} \sum_{j \in s} c_{sij} \hat{G}_i \hat{G}_j\right) \tag{12.5.5}$$

where $\hat{G}_i = Y_i - \hat{\bar{\bar{Y}}}_{cs} M_i$, $c_{si} = \alpha_i/\pi_i$, $c_{sij} = \alpha_{ij}/\pi_{ij}$, $\pi_i =$ inclusion probability for the $i$th unit, and $\pi_{ij} =$ inclusion probability for the $i$th and $j(\neq i)$th unit.

## 12.5.1 Examples

### 12.5.1.1 Arbitrary Sampling Design

Consider a fixed effective size $n$ sampling design with $\pi_i > 0$. For $b_{si} = 1/\pi_i$, we have

$$\hat{\bar{\bar{Y}}}_{Rc}(ht) = \left(\sum_{i\in s} Y_i/\pi_i\right) \Big/ \left(\sum_{i\in s} M_i/\pi_i\right),$$

$$B\left(\hat{\bar{\bar{Y}}}_{Rc}\right) = -\frac{1}{\left(N\overline{M}\right)^2}\frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{M_i}{\pi_i} - \frac{M_j}{\pi_j}\right)\left(\frac{G_i}{\pi_i} - \frac{G_j}{\pi_j}\right)$$

$$M\left(\hat{\bar{\bar{Y}}}_{Rc}\right) = \frac{1}{\left(N\overline{M}\right)^2}\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{G_i}{\pi_i} - \frac{G_j}{\pi_j}\right)^2,$$

$$\widehat{B}\left(\hat{\bar{\bar{Y}}}_{Rc}\right) = -\frac{1}{\widehat{M}_{cs}^2}\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{M_i}{\pi_i} - \frac{M_j}{\pi_j}\right)\left(\frac{\widehat{G}_i(ht)}{\pi_i} - \frac{\widehat{G}_j(ht)}{\pi_j}\right)$$

and

$$\widehat{M}\left(\hat{\bar{\bar{Y}}}_{Rc}\right) = \frac{1}{\left(\widehat{M}_{cs}\right)^2}\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{\widehat{G}_i(ht)}{\pi_i} - \frac{\widehat{G}_j(ht)}{\pi_j}\right)^2$$

where $\widehat{M}_{cs} = \sum_{i\in s} M_i/\pi_i$ and $\widehat{G}_i(ht) = Y_i - \hat{\bar{\bar{Y}}}_{Rc}(ht)M_i$.

### 12.5.1.2 Simple Random Sampling Without Replacement

For SRSWOR sampling $b_{si} = N/n$ and hence

$$\hat{\bar{\bar{Y}}}_{Rc}(wor) = \sum_{i\in s} Y_i \Big/ \sum_{i\in s} M_i,$$

$$B\left[\hat{\bar{\bar{Y}}}_{Rc}(wor)\right] = -\frac{(1-f)}{n\overline{M}^2}S_{GM}$$

$$M\left[\hat{\bar{\bar{Y}}}_{Rc}(wor)\right] = \frac{(1-f)}{n\overline{M}^2}S_G^2$$

$$\widehat{B}\left[\hat{\bar{\bar{Y}}}_{Rc}(wor)\right] = -\frac{(1-f)}{n\overline{M}_s^2}s_{\hat{G}M}$$

and

$$\widehat{M}\left[\widehat{\overline{\overline{Y}}}_{Rc}(wor)\right] = \frac{(1-f)}{n\widehat{\overline{M}}_s^2} s_{\hat{G}}^2$$

where $f = n/N$, $\widehat{\overline{M}}_s = \sum_{i\in s} M_i/n$, $S_G^2 = \sum_{i\in U} \left(G_i - \overline{G}\right)^2/(N-1)$,

$\overline{G} = \sum_{i\in U} G_i/N = 0$, $S_{GM} = \sum_{i=1}^{N} G_i\left(M_i - \overline{M}\right)/(N-1)$,

$s_{\hat{G}}^2 = \sum_{i\in s}\left(\widehat{G}_{iwor} - \widehat{\overline{G}}_s\right)^2/(n-1)$, $\widehat{G}_{iwor} = Y_i - \widehat{\overline{\overline{Y}}}_{Rc}(wor)M_i$, and

$\widehat{\overline{G}}_s = \sum_{i\in s} \widehat{G}_{iwor}/n$.

## 12.6 EXERCISES

**12.6.1** A sample $s$ of $n$ clusters is selected from a population of $N$ clusters of unequal sizes by SRSWOR method. Show that the estimator $t = \frac{1}{n}\sum_{i\in s} \overline{Y}_i$ is biased estimator for the mean per unit $\overline{\overline{Y}}$. Find the expressions of (i) bias, (ii) mean square error, and (iii) their unbiased estimators

**12.6.2** Let $s$ be a sample of size $n$ clusters is selected from a population of $N$ clusters by SRSWOR method and all the clusters are of equal size $M$. Furthermore, let $P_i$ and $P$ be the proportions of units that possess a certain attribute in the $i$th cluster and entire population, respectively. Show that (i) $\widehat{P} = \sum_{i\in s} \widehat{P}_i/n$ is an unbiased estimator for $P$ and (ii) $V\left(\widehat{P}\right) = \frac{(1-f)}{n}\sum_{i=1}^{N}(P_i - P)^2/(N-1)$.

**12.6.3** Let $NM$ units of a population be grouped at random into $N$ clusters each of size $M$ and let $s$ be a sample of $n$ clusters selected from $N$ clusters by SRSWOR method. Show that (i) $t = \frac{1}{n}\sum_{i\in s} \overline{Y}_i$ is an unbiased estimator of the mean per unit $\overline{\overline{Y}}$ and (ii) $V(t) = (1-f)S_y^2/(nM)$, where $S_y^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \overline{\overline{Y}}\right)^2/(N-1)$ and $f = n/N$.

**12.6.4** Let a finite population of $M_0$ units be divided into $N$ clusters so that the $i$th cluster contains $M_i$ units $\left(\sum_{i=1}^{N} M_i = M_0\right)$. A sample of $m$

units is selected by SRSWOR method from the $M_0$ units and let $Y_i$ be the total of the study variable $y$ that belongs to the $i$th cluster ($Y_i = 0$ if none of the units fall into the $i$th cluster). Let a sample $s$ of $n$ clusters be selected from the $N$ clusters by SRSWOR method. Show that $t = \frac{N}{nm} \sum_{i \in s} Y_i$ is an unbiased estimator of the population mean $\overline{\overline{Y}}$ and obtain the variance of $t$ (Ghosh, 1963).

**12.6.5** A sample of 10 households is selected from a locality of 50 households by SRSWOR method. The household size, number of earning members, and the total monthly income of the selected households are given in the following table.

| Serial no. of households | Household size | Number of earning members | Total income in $ |
|---|---|---|---|
| 1 | 5 | 2 | 7000 |
| 2 | 3 | 1 | 5000 |
| 3 | 4 | 2 | 8000 |
| 4 | 2 | 1 | 4000 |
| 5 | 4 | 1 | 5000 |
| 6 | 1 | 1 | 3000 |
| 7 | 4 | 2 | 7000 |
| 8 | 5 | 1 | 3000 |
| 9 | 4 | 2 | 7000 |
| 10 | 2 | 2 | 6000 |

Estimate the average household income and proportion of earning members per household of the locality along with their standard errors.

**12.6.6** The following table gives the yearly production of apples in 10 orchards selected at random from 60 orchards in a certain block.

| Serial no. of orchards | No. of trees | Yield of orchards in Kg |
|---|---|---|
| 1 | 6 | 200, 150, 720, 420, 400, 200 |
| 2 | 3 | 250, 200, 200 |
| 3 | 5 | 320, 470, 125, 375, 120 |
| 4 | 2 | 250, 400 |
| 5 | 4 | 175, 125, 195, 370 |
| 6 | 1 | 300 |
| 7 | 4 | 180, 120, 320, 170 |
| 8 | 5 | 250, 175, 300, 420, 180 |
| 9 | 4 | 370, 480, 200, 150 |
| 10 | 2 | 400, 250 |

(i) Estimate the total production of apples of the block and also the average production of apples per orchard. Give unbiased estimates of the variances of the proposed estimators.
(ii) Find a 95% confidence interval of the average production of apple per tree.

**12.6.7** To estimate availability of doctors in a locality, 5 clinics were selected from 100 clinics by SRSWOR method. The results are given in the following table.

| Serial no. of clinic | No. of doctors | No. of patients treated per day |
|---|---|---|
| 1 | 3 | 20, 15, 40 |
| 2 | 3 | 25, 20, 20 |
| 3 | 3 | 20, 27, 12 |
| 4 | 3 | 25, 40, 20 |
| 5 | 4 | 15, 25, 20 |

(i) Estimate the intracluster correlation coefficients between doctors within clinics.
(ii) Estimate the average number of patients treated per clinic when the number of doctors in each of the clinic is fixed at 3. Estimate the efficiency of cluster sampling over SRSWOR sampling for estimating the average number of patients treated by a doctor.