CHAPTER 13

Multistage Sampling

13.1 INTRODUCTION

In cluster sampling, the entire population is divided into a number (N) of mutually exclusive and exhaustive groups called clusters. Each cluster consists of several ultimate units. A sample of $n(<N)$ clusters is chosen by some suitable sampling scheme, and all the ultimate units of the selected clusters are surveyed. In two-stage sampling, "clusters" are termed as "first-stage units (fsu's)" or "primary sampling units." The ultimate units in a cluster are called "second-stage units (ssu's)." In two-stage sampling, a sample s of n fsu's is selected by a sampling procedure. Then from each of the selected fsu's, a sample of ssu's of suitable sizes are selected by a suitable sampling procedure. Clearly, two-stage sampling reduces to a cluster sampling if all of the sampled fsu's are completely enumerated. For example, suppose we want to select 5000 households from KwaZulu-Natal (KZN) to determine their average income. In this situation we can divide KZN into a number of municipalities, which may be regarded as fsu's. Then we may select a sample of fsu's, and from each of the selected fsu's (municipalities) a suitable number of ssu's (households) may be selected by a suitable sampling scheme. If a sample of ultimate units is selected by more than two stages, then the sampling procedure is called multistage sampling. For example, we may divide South Africa into different provinces, then each province is divided again into municipalities, and municipalities may be divided into households. So, in three-stage sampling we first select a sample of provinces, then from each of the selected province, select a sample of municipalities, and finally from the selected municipalities, households may be selected. The multistage sampling is a compromise between a cluster sampling and a unistage sampling (units are directly selected from the population). Hence it is much cheaper and more convenient to draw a sample in a two-stage sampling than a unistage sampling procedure, but more expensive than a cluster sampling. With regard to efficiency, multistage sampling is generally more efficient than

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00013-3

cluster sampling but less efficient than unistage sampling. Multistage sampling is widely used in most large-scale surveys because the sampling frame of the ultimate units is not generally available in practice. For the selection of ssu's, only the sampling frames of the selected fsu's are required. Similarly, for the selection of the third-stage units (tsu's), only the sampling frames of the selected ssu's are required.

13.2 TWO-STAGE SAMPLING SCHEME

Let us consider a population U consisting of N fsu's and the ith fsu consists of M_i ssu's and the total number of ssu's in the population $M_0 = \sum_{i \in U} M_i$. Let y_{ij} be the value of the character y, under study for the jth ssu of the ith fsu, $j = 1,\ldots, M_i$; $i = 1,\ldots, N$ and $Y_i = \sum_{j=1}^{M_i} y_{ij}$ = total of the y-values of the ssu's that belong to the ith fsu. $\overline{Y}_i = Y_i/M_i$ = mean of the ssu's that belong to the ith fsu. The population total for all the ssu's of the population is $Y = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} Y_i$ and the population mean per unit is $\overline{\overline{Y}} = Y/M_0$.

Let a sample s of size n fsu's be selected from the N fsu's of the population U with probability $p(s)$ by some suitable sampling design. Let π_i and π_{ij} be the inclusion probabilities for the ith and i and jth ($i \neq j$) fsu's. If the ith fsu is selected in the sample s, i.e., if $i \in s$, then from each of the ith ($i \in s$) fsu, a subsample s_i of predetermined size m_i ssu's is selected independently by following a suitable sampling design.

13.3 ESTIMATION OF THE POPULATION TOTAL AND VARIANCE

Let $\widehat{Y}_i$ be an unbiased estimator for the total Y_i based on the sample s_i ($i \in s$). Then we may consider an unbiased estimator for the population total Y as

$$\widehat{Y}_{ts} = \sum_{i \in s} b_{si} \widehat{Y}_i \tag{13.3.1}$$

The coefficients b_{si}'s are known and satisfy the unbiasedness condition $\sum_{s \supset i} b_{si} p(s) = 1$.

Theorem 13.3.1

(i) $E\left(\widehat{Y}_{ts}\right) = Y$

(ii) $$V\left(\widehat{Y}_{ts}\right) = \sum_{i\in U} \alpha_i Y_i^2 + \sum_{i\neq} \sum_{j\in U} \alpha_{ij} Y_i Y_j + \sum_{i\in U} \beta_i \sigma_i^2 \tag{13.3.2}$$

where $\alpha_i = \sum_{s\supset i} b_{si}^2 p(s) - 1$, $\alpha_{ij} = \sum_{s\supset i,j} b_{si} b_{sj} p(s) - 1$, $\beta_i = \alpha_i + 1$, and $\sigma_i^2 = V\left(\widehat{Y}_i \middle| s\right)$

Proof

Let E_1 and V_1 be the expectation and variance operators for the selection of the sample s of the fsu's, and let $E_2(\cdot|s)$ and $V_2(\cdot|s)$ be the conditional expectation and variance for a given s. Then,

(i) $$\begin{aligned} E\left(\widehat{Y}_{ts}\right) &= E_1\left[\sum_{i\in s} b_{si} E_2\left(\widehat{Y}_i \middle| s\right)\right] \\ &= E_1\left(\sum_{i\in s} b_{si} Y_i\right) \\ &= Y \end{aligned}$$

(ii) $$\begin{aligned} V\left(\widehat{Y}_{ts}\right) &= V_1\left[E_2\left(\sum_{i\in s} b_{si} \widehat{Y}_i \middle| s\right)\right] + E_1\left[V_2\left(\sum_{i\in s} b_{si} \widehat{Y}_i \middle| s\right)\right] \\ &= V_1\left[\sum_{i\in s} b_{si} E_2\left(\widehat{Y}_i \middle| s\right)\right] + E_1\left[\sum_{i\in s} b_{si}^2 V_2\left(\widehat{Y}_i \middle| s\right)\right] \\ &= V_1\left(\sum_{i\in s} b_{si} Y_i\right) + E_1\left(\sum_{i\in s} b_{si}^2 \sigma_i^2\right) \end{aligned}$$

Now

$$\begin{aligned} V_1\left(\sum_{i\in s} b_{si} Y_i\right) &= \sum_{i\in U} Y_i^2 \sum_{s\supset i} b_{si}^2 p(s) + \sum_{i\neq} \sum_{j\in U} Y_i Y_j \sum_{s\supset i,j} b_{si} b_{sj} p(s) - Y^2 \\ &= \sum_{i\in U} \alpha_i Y_i^2 + \sum_{i\neq} \sum_{j\in U} \alpha_{ij} Y_i Y_j \end{aligned}$$

and

$$E_1\left(\sum_{i\in s} b_{si}^2\sigma_i^2\right) = \sum_{i\in U}\sigma_i^2\sum_{s\supset i} b_{si}^2 p(s) = \sum_{i\in U}\beta_i\sigma_i^2 \text{ Q.E.D.}$$

Theorem 13.3.2

Let $Q_s(Y) = \sum_{i\in s} c_{si}Y_i^2 + \sum_{i\neq}\sum_{j\in s} c_{sij}Y_iY_j$ be an unbiased estimator of $V_1\left(\sum_{i\in s} b_{si}Y_i\right)$, satisfying $E_1(Q_s(Y)) = V_1\left(\sum_{i\in s} b_{si}Y_i\right)$, then an unbiased estimator of $V\left(\widehat{Y}_{ts}\right)$ is

$$\widehat{V}\left(\widehat{Y}_{ts}\right) = Q_s\left(\widehat{Y}\right) + \widehat{\mathbf{\Phi}}_s$$

where $Q_s\left(\widehat{Y}\right) = \sum_{i\in s} c_{si}\widehat{Y}_i^2 + \sum_{i\neq}\sum_{j\in s} c_{sij}\widehat{Y}_i\widehat{Y}_j$, $\widehat{\mathbf{\Phi}}_s$ is an unbiased estimator of $\sum_{i\in U}\sigma_i^2$; c_{si} and c_{sij} are suitably chosen constants.

Proof

Since $Q_s(Y)$ is an unbiased estimator of $V_1\left(\sum_{i\in s} b_{si}Y_i\right)$, c_{si} and c_{sij} must satisfy

$$\sum_{s\supset i} c_{si}p(s) = \alpha_i \text{ and } \sum_{s\supset i,j} c_{sij}p(s) = \alpha_{ij} \tag{13.3.3}$$

Now noting

$$\begin{aligned}
E\left[Q_s\left(\widehat{Y}\right)\right] &= E_1\left[E_2\left(\sum_{i\in s} c_{si}\widehat{Y}_i^2 + \sum_{i\neq}\sum_{j\in s} c_{sij}\widehat{Y}_i\widehat{Y}_j\right)\right] \\
&= E_1\left[\sum_{i\in s} c_{si}E_2\left(\widehat{Y}_i^2\middle|s\right) + \sum_{i\neq}\sum_{j\in s} c_{sij}E_2\left(\widehat{Y}_i\middle|s\right)E_2\left(\widehat{Y}_j\middle|s\right)\right] \\
&= E_1\left[\sum_{i\in s} c_{si}\left(Y_i^2+\sigma_i^2\right) + \sum_{i\neq}\sum_{j\in s} c_{sij}Y_iY_j\right] \\
&= E_1[Q_s(Y)] + \sum_{i\in U}\alpha_i\sigma_i^2 \\
&= V_1\left(\sum_{i\in s} b_{si}Y_i\right) + \sum_{i\in U}\alpha_i\sigma_i^2
\end{aligned}$$

we get

$$
\begin{aligned}
E\left[\widehat{V}\left(\widehat{Y}_{ts}\right)\right] &= E\left[Q_s\left(\widehat{Y}\right)\right] + E\left(\widehat{\Phi}_s\right) \\
&= V_1\left(\sum_{i\in s} b_{si}Y_i\right) + \sum_{i\in U}\alpha_i\sigma_i^2 + \sum_{i\in U}\sigma_i^2 \\
&= \sum_{i\in U}\alpha_i Y_i^2 + \sum_{i\neq}\sum_{j\in U}\alpha_{ij}Y_iY_j + \sum_{i\in U}\beta_i\sigma_i^2 \\
&= V\left(\widehat{Y}\right) \text{ Q.E.D.}
\end{aligned}
$$

The appropriate choices of c_{si}, c_{sij}, and $\widehat{\Phi}_s$ depend on the sampling design under considerations, but obvious choices are $c_{si} = \alpha_i/\pi_i$, $c_{sij} = \alpha_{ij}/\pi_{ij}$, and $\widehat{\Phi}_s = \sum_{i\in s}\widehat{\sigma}_i^2/\pi_i$ or $\sum_{i\in s} b_{si}\widehat{\sigma}_i^2$ where $\widehat{\sigma}_i^2$ is the conditionally unbiased estimator of σ_i^2, which satisfies $E_2\left(\widehat{\sigma}_i^2\right) = \sigma_i^2$. Detailed discussions of various choices of c_{si}, c_{sij}, and $\widehat{\Phi}_s$ have been given by Chaudhuri and Arnab (1982b).

13.3.1 First-Stage Arbitrary Sampling Designs and Second-Stage Simple Random Sampling Without Replacement

Let the sample s be selected by an arbitrary fixed effective size (n) sampling design with positive inclusion probabilities π_i and π_{ij} for the ith, and ith and $j(\neq i)$ units, respectively. Each of the selected fsu's in s be subsampled independently by simple random sampling without replacement (SRSWOR) method. Let

$$\widehat{Y}_i = M_i\overline{y}(s_i) \tag{13.3.4}$$

with $\overline{y}(s_i) = \sum_{j\in s_i} y_{ij}/m_i$.

Then,

$$\sigma_i^2 = V_2\left(\widehat{Y}_i\middle|s\right) = M_i^2(1-f_i)S_{yi}^2/m_i \tag{13.3.5}$$

where $f_i = m_i/M_i$, $S_{yi}^2 = \sum_{j=1}^{M_i}\left(y_{ij}-\overline{Y}_i\right)^2\Big/(M_i-1)$ and $\overline{Y}_i = Y_i/M_i$.

Now using the Theorems 13.3.1 and 13.3.2, we note that

$$\widehat{Y}_{ts}(ht) = \sum_{i\in s}\widehat{Y}_i/\pi_i = \sum_{i\in s} M_i\overline{y}(s_i)/\pi_i \tag{13.3.6}$$

is an unbiased estimator of the population total Y with variance

$$\begin{aligned} V\left[\widehat{Y}_{ts}(ht)\right] &= V_1\left(\sum_{i\in s} Y_i/\pi_i\right) + E_1\left(\sum_{i\in s}\sigma_i^2/\pi_i^2\right) \\ &= \frac{1}{2}\sum_{i\neq}\sum_{j\in U}\Delta_{ij}\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2 + \sum_{i\in U}\frac{M_i^2(1-f_i)S_{yi}^2}{m_i\pi_i} \end{aligned} \quad (13.3.7)$$

where $\Delta_{ij} = \pi_i\pi_j - \pi_{ij}$, $f_i = m_i/M_i$ and $S_{yi}^2 = \sum_{j=1}^{M_i}\left(y_{ij} - \overline{Y}_i\right)^2$.

Further noting that $Q_s(Y) = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2$ and $\widehat{\sigma}_i^2 = M_i^2(1-f_i)s_{yi}^2\Big/m_i$ are unbiased estimators for $V_1\left(\sum_{i\in s}Y_i/\pi_i\right) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}\Delta_{ij}\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2$ and σ_i^2, respectively, we find an unbiased estimator of $V\left(\widehat{Y}_{ms}(ht)\right)$ as

$$\widehat{V}\left(\widehat{Y}_{ts}(ht)\right) = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{\widehat{Y}_i}{\pi_i}-\frac{\widehat{Y}_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{M_i^2(1-f_i)s_{yi}^2}{m_i\pi_i} \quad (13.3.8)$$

where $s_{yi}^2 = \sum_{j\in s_i}\left\{y_{ij} - \overline{y}(s_i)\right\}^2/(m_i-1)$ and $f_i = m_i/M_i$.

13.3.2 Simple Random Sampling Without Replacement Both the Stages

For an SRSWOR sampling, $\pi_i = n/N$, $\pi_{ij} = n(n-1)/\{N(N-1)\}$, and Eqs. (13.3.6)−(13.3.8), reduce to

$$\widehat{Y}_{ts}(wor) = \frac{N}{n}\sum_{i\in s}M_i\overline{y}(s_i) \quad (13.3.9)$$

$$V\left(\widehat{Y}_{ts}(wor)\right) = N^2(1-f)\frac{S_b^2}{n} + \frac{N}{n}\sum_{i\in U}\frac{M_i^2(1-f_i)S_{yi}^2}{m_i} \quad (13.3.10)$$

and

$$\widehat{V}\left(\widehat{Y}_{ts}(wor)\right) = N^2(1-f)\frac{s_b^2}{n} + \frac{N}{n}\sum_{i\in s}\frac{M_i^2(1-f_i)s_{yi}^2}{m_i} \quad (13.3.11)$$

where $S_b^2 = \sum_{i\in U}\left(Y_i - \overline{Y}\right)^2/(N-1)$, $s_b^2 = \sum_{i\in s}\left(\widehat{Y}_i - \widehat{\overline{Y}}_s\right)^2\Big/(n-1)$, $f = n/N$, $\overline{Y} = \sum_{i\in U}Y_i/N$ and $\widehat{\overline{Y}}_s = \sum_{i\in s}\widehat{Y}_i\big/n = \sum_{i\in s}M_i\overline{y}(s_i)/n$.

13.3.3 First-Stage Rao–Hartley–Cochran and Second-Stage Simple Random Sampling Without Replacement

Here we suppose that the sample s of n fsu's is selected by Rao–Hartley–Cochran (1962) method of sampling using normed size measure p_i attached to the ith unit and P_i is the sum of p_j's for the group containing the ith unit that was formed in selecting the sample s (the detailed selection procedure has been described in Section 5.6). The ssu's are selected independently from the selected fsu's by SRSWOR method. For this sampling scheme, an unbiased estimator for the total Y is given by

$$\widehat{Y}_{ts}(rhc) = \sum_{i\in s}\frac{\widehat{Y}_i}{p_i}P_i \tag{13.3.12}$$

where $\widehat{Y}_i$ is given in Eq. (13.3.4).

The variance of $\widehat{Y}$ is given by

$$V\left(\widehat{Y}_{ts}(rhc)\right) = V_1\left(\sum_{i\in s}\frac{Y_i}{p_i}P_i\right) + E_1\left(\sum_{i\in s}\frac{\sigma_i^2}{p_i^2}P_i^2\right) \tag{13.3.13}$$

Now using Eq. (5.6.7) and Arnab (2004a,b), we get

$$V_1\left(\sum_{i\in s}\frac{Y_i}{p_i}P_i\right) = \frac{N-n}{n(N-1)}\sum_{i\in U}p_i\left(\frac{Y_i}{p_i}-Y\right)^2 \tag{13.3.14}$$

and

$$E_1\left(\sum_{i\in s}\frac{\sigma_i^2}{p_i^2}P_i^2\right) = \frac{N(n-1)}{n(N-1)}\sum_{i\in U}\sigma_i^2 + \frac{N-n}{n(N-1)}\sum_{i\in U}\frac{\sigma_i^2}{p_i} \tag{13.3.15}$$

On substituting Eqs. (13.3.14) and (13.3.15) in Eq. (13.3.13), we get

$$V\left(\widehat{Y}_{ts}(rhc)\right) = \frac{N-n}{n(N-1)}\sum_{i\in U}p_i\left(\frac{Y_i}{p_i}-Y\right)^2 + \frac{N(n-1)}{n(N-1)}\sum_{i\in U}\sigma_i^2 + \frac{N-n}{n(N-1)}\sum_{i\in U}\frac{\sigma_i^2}{p_i} \tag{13.3.16}$$

Noting $Q_s(Y) = \frac{N-n}{N(n-1)} \sum_{i\in s} P_i \left(\frac{Y_i}{p_i} - \sum_{i\in s} \frac{Y_i}{p_i} P_i \right)^2$ is an unbiased estimator of $V_1\left(\sum_{i\in s} \frac{Y_i}{p_i} P_i \right)$ and $\widehat{\Phi}_s = \sum_{i\in s} \frac{\widehat{\sigma}_i^2}{p_i} P_i$ is an unbiased estimator Φ with $\widehat{\sigma}_i^2 = M_i^2(1-f_i)s_{yi}^2 \big/ m_i$, we get an unbiased estimator of $V\left(\widehat{Y}\right)$ as

$$\begin{aligned}\widehat{V}\left(\widehat{Y}_{ts}(rhc)\right) &= Q_s\left(\widehat{Y}\right) + \widehat{\Phi}_s \\ &= \frac{N-n}{N(n-1)} \sum_{i\in s} P_i \left(\frac{\widehat{Y}_i}{p_i} - \widehat{Y} \right)^2 + \sum_{i\in s} \frac{M_i^2(1-f_i)s_{yi}^2}{m_i p_i} P_i\end{aligned} \tag{13.3.17}$$

13.4 FIRST-STAGE UNITS ARE SELECTED BY PPSWR SAMPLING SCHEME

Consider the situation when the sample s of n fsu's are selected by probability proportional to size with replacement (PPSWR) method using the normed size measure p_i attached to the ith unit. If the ith fsu is selected $n_i(s)$ times in the sample, then $n_i(s)$-independent subsamples each of size m_i are selected from the ith fsu using the same sampling design. In this situation, we consider the following Hansen and Hurwitz (1943) type estimator for the total Y as follows:

$$\widehat{Y}_{ts}(hh) = \frac{1}{n} \sum_{r=1}^{n} \frac{\widehat{Y}(r)}{p(r)} \tag{13.4.1}$$

Here $\widehat{Y}(r)$ is an unbiased estimator of the fsu total that was selected at the rth draw with probability $p(r)$, i.e., if the jth fsu is selected at the rth draw, then $\widehat{Y}(r) = \widehat{Y}_j$ and $p(r) = p_j$. In this case

$$E_2\left(\frac{\widehat{Y}(r)}{p(r)} \right) = \frac{Y(r)}{p(r)} \quad \text{and} \quad V_2\left(\frac{\widehat{Y}(r)}{p(r)} \right) = \frac{\sigma^2(r)}{\{p(r)\}^2} \tag{13.4.2}$$

where $Y(r)$ = fsu total of the fsu that is selected at the rth draw and $\sigma^2(r)$ variance of $\widehat{Y}(r)$. Hence

$$E\left(\frac{\widehat{Y}(r)}{p(r)} \right) = E_1\left[E_2\left(\frac{\widehat{Y}(r)}{p(r)} \right) \right] = E_1\left(\frac{Y(r)}{p(r)} \right) = \sum_{i\in U} \frac{Y_i}{p_i} p_i = Y \tag{13.4.3}$$

and

$$V\left(\frac{\widehat{Y}(r)}{p(r)}\right) = E_1\left[V_2\left(\frac{\widehat{Y}(r)}{p(r)}\right)\right] + V_1\left[E_2\left(\frac{\widehat{Y}(r)}{p(r)}\right)\right]$$

$$= E_1\left(\frac{\sigma^2(r)}{p^2(r)}\right) + V_1\left(\frac{Y(r)}{p(r)}\right) \tag{13.4.4}$$

$$= \sum_{i\in U}\frac{\sigma_i^2}{p_i} + \sum_{i\in U} p_i\left(\frac{Y_i}{p_i} - Y\right)^2 = W^2(\text{say})$$

Since $\dfrac{\widehat{Y}(r)}{p(r)}$'s are independently and identically distributed with common mean Y and common variance W^2, we get the following theorem.

Theorem 13.4.1

(i) $E\left[\widehat{Y}_{ts}(hh)\right] = E\left(\dfrac{1}{n}\displaystyle\sum_{r=1}^{n}\dfrac{\widehat{Y}(r)}{p(r)}\right) = Y$

(ii) $V\left[\widehat{Y}_{ts}(hh)\right] = \dfrac{W^2}{n} = \dfrac{1}{n}\left[\displaystyle\sum_{i\in U}\dfrac{\sigma_i^2}{p_i} + \sum_{i\in U} p_i\left(\dfrac{Y_i}{p_i} - Y\right)^2\right]$

(iii) An unbiased estimator of $V\left[\widehat{Y}_{ts}(hh)\right]$ is

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

13.4.1 Simple Random Sampling With Replacement

In particular, if the fsu's are selected by simple random sampling with replacement (SRSWR) sampling scheme and if the selected fsu's are subsampled independently by SRSWOR method, we then get $\widehat{Y}(r) = M(r)\overline{y}(r)$, where $M(r)$ and $\overline{y}(r)$ are the total number of ssu's and the sample mean of the ssu's of the fsu that was selected at the rth draw, respectively. In this case, $E_2\left\{\widehat{Y}(r)\right\} = Y(r)$ and $V_2\left\{\widehat{Y}(r)\right\} = \sigma^2(r)$. Thus if the rth draw produces ith unit, we have $E_2\left\{\widehat{Y}(r)\right\} = Y_i$ and $V_2\left\{\widehat{Y}(r)\right\} = \sigma^2(i) = \dfrac{M_i^2(1-f_i)S_{yi}^2}{m_i}$ with probability $1/N$. Hence, under SRSWR sampling we have the following theorem.

Theorem 13.4.2

Under SRSWR sampling,

(i) $\widehat{Y}_{ts}(wr) = \frac{N}{n}\sum_{r=1}^{n} M(r)\overline{y}(r)$ is an unbiased estimator for Y

(ii) $V\left[\widehat{Y}_{ts}(wr)\right] = \frac{N^2}{n}\left[\sum_{i\in U}\frac{M_i^2(1-f_i)S_{yi}^2}{m_i} + (N-1)S_b^2\right]$

(iii) $\widehat{V}\left[\widehat{Y}_{ts}(wr)\right] = \frac{N^2}{n(n-1)}\sum_{r=1}^{n}\left(M(r)\overline{y}(r) - \frac{1}{n}\sum_{r=1}^{n} M(r)\overline{y}(r)\right)^2$ is an unbiased estimator of $V\left[\widehat{Y}_{ts}(wr)\right]$.

Remark 13.4.1

So far, we have considered σ_i^2 as independent of s, the sample of fsu's. Rao (1975) considered a situation where σ_i^2 may depend on s; i.e., $V_2\left(\widehat{Y}_i\right) = \sigma_{si}^2$. In this situation, following Theorems 13.4.1 and 13.4.2, the variance of $\widehat{Y}_{ts}$ and its unbiased estimator become,

$$V\left(\widehat{Y}_{ts}\right) = \sum_{i\in U}\alpha_i Y_i^2 + \sum_{i\neq}\sum_{j\in U}\alpha_{ij}Y_iY_j + \sum_{s}\left(\sum_{i\in s} b_{si}^2\sigma_{si}^2\right)p(s) \qquad (13.4.5)$$

and

$$\widehat{V}\left(\widehat{Y}_{ts}\right) = \sum_{i\in s} c_{si}\widehat{Y}_i^2 + \sum_{i\neq}\sum_{j\in s} c_{sij}\widehat{Y}_i\widehat{Y}_j + \sum_{i\in s}\left(b_{si}^2 - c_{si}\right)\widehat{\sigma}_{si}^2 \qquad (13.4.6)$$

respectively, where $\widehat{\sigma}_{si}^2$ is an unbiased estimator of σ_{si}^2 that satisfies $E_2\left(\widehat{\sigma}_{si}^2\right) = \sigma_{si}^2$.

13.4.2 Raj Estimator for Multi-Stage Sampling

Suppose that from a population of N fsu's, a sample s of n fsu's is selected by the PPSWR method using p_i as the normed size measure for the ith unit. If the ith fsu appears in s, $n_i(s)$ times, a subsample of size $n_i(s)$ m_i ssu's is selected from the ith fsu by SRSWOR method, assuming $m_i n \leq M_i$. We then have the following theorem according to Raj (1968).

Theorem 13.4.3

(i) $\widehat{Y}^*_{ts}(hh) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{N}\frac{n_i(s)}{p_i}\widehat{Y}_i$ is an unbiased estimator for Y

(ii) $$V\left[\widehat{Y}^*_{ts}(hh)\right] = \frac{1}{n}\left[\sum_{i\in U}p_i\left(\frac{Y_i}{p_i} - Y\right)^2 + \sum_{i\in U}\left\{\frac{1}{p_i}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) - \frac{n-1}{M_i}\right\}M_i^2 S_{yi}^2\right]$$

(iii) $$\widehat{V}\left[\widehat{Y}^*_{ts}(hh)\right] = \frac{\sum_i n_i(s)\left(\dfrac{\widehat{Y}_i}{p_i} - \widehat{Y}^*_{ts}(hh)\right)^2}{n(n-1)} + \sum_i \frac{n_i(s)(n_i(s)-1)\widehat{\sigma}_{si}^2}{n(n-1)p_i^2}$$

where $$\widehat{\sigma}_{si}^2 = M_i^2\left(\frac{1}{n_i(s)m_i} - \frac{1}{M_i}\right)\frac{\sum_{j\in s_i}\left\{y_{ij} - \overline{y}(s_i)\right\}^2}{n_i(s)m_i - 1}$$

Proof

(i) $$E\left[\widehat{Y}^*_{ts}(hh)\right] = \frac{1}{n}E_1\left[\sum_{i=1}^{N}\frac{n_i(s)}{p_i}\left\{E_2\left(\widehat{Y}_i\right)\right\}\right] = \frac{1}{n}E_1\sum_{i=1}^{N}\frac{n_i(s)}{p_i}Y_i = Y$$

(ii) $$V\left[\widehat{Y}^*_{ts}(hh)\right] = V_1\left(\frac{1}{n}\sum_{i=1}^{N}\frac{n_i(s)}{p_i}Y_i\right) + E_1\left(\frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{n_i(s)}{p_i}\right)^2\sigma_i^2(s)\right)$$

where $$\sigma_i^2(s) = M_i^2\left(\frac{1}{n_i(s)m_i} - \frac{1}{M_i}\right)S_{yi}^2$$

Now noting $V\left(\dfrac{n_i(s)}{np_i}\right) = \dfrac{(1-p_i)}{np_i}$ and $Cov\left(\dfrac{n_i(s)}{np_i}, \dfrac{n_j(s)}{np_i}\right) = -\dfrac{1}{n}$, we obtain

$$V\left[\widehat{Y}^*_{ts}(hh)\right] = \frac{1}{n}\left[\sum_{i\in U}p_i\left(\frac{Y_i}{p_i} - Y\right)^2 + \sum_{i\in U}\left\{\frac{1}{p_i}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) - \frac{n-1}{M_i}\right\}M_i^2 S_{yi}^2\right]$$

(iii) $n(n-1)E\left[\widehat{V}\left\{\widehat{Y}^{*}_{ts}(hh)\right\}\right]$

$$= E_1\sum_i \frac{n_i(s)\left\{V_2\left(\widehat{Y}_i|s\right)+Y_i^2\right\}}{p_i^2} - n\left[V\left\{\widehat{Y}^{*}_{ts}(hh)\right\}+Y^2\right]$$

$$+ E_1\sum_i \frac{n_i(s)\{n_i(s)-1\}E\left(\widehat{\sigma}^2_{si}|s\right)}{p_i^2}$$

$$= E_1\sum_i \frac{n_i(s)Y_i^2}{p_i^2} - n\left(V\left\{\widehat{Y}^{*}_{ts}(hh)\right\}+Y^2\right) + E_1\sum_i \frac{\{n_i(s)\}^2\sigma^2_{si}}{p_i^2}$$

$$= n\sum_i p_i\left(\frac{Y_i}{p_i}-Y\right)^2 - nV\left\{\widehat{Y}^{*}_{ts}(hh)\right\} + E_1\sum_i \frac{\{n_i(s)\}^2}{p_i^2}\left(\frac{1}{m_i n_i(s)}-\frac{1}{M_i}\right)S^2_{yi}$$

$$= n\sum_i p_i\left(\frac{Y_i}{p_i}-Y\right)^2 - nV\left\{\widehat{Y}^{*}_{ts}(hh)\right\} + n\sum_i \frac{1}{p_i}\left(\frac{1}{m_i}-\frac{1+(n-1)p_i}{M_i}\right)M_i^2 S^2_{yi}$$

$$= n(n-1)V\left[\widehat{Y}^{*}_{ts}(hh)\right]$$

i.e., $E\left[\widehat{V}\left\{\widehat{Y}^{*}_{ts}(hh)\right\}\right] = V\left[\widehat{Y}^{*}_{ts}(hh)\right]$

Remark 13.4.2

It is important to note from Theorems 13.3.2 and 13.4.1 that if the fsu's are selected by without replacement sampling schemes or if the estimator of the total $\widehat{Y}_{ts}$ is based on distinct units for a with replacement sampling scheme, then the unbiased estimators for the variance of $\widehat{Y}_{ts}$ involve $\widehat{\sigma}^2_i$, the unbiased estimator of the variance of $\widehat{Y}_i$. But if the fsu's are selected by PPSWR (or SRSWR), then the variance of $\widehat{Y}_{ts}$ can be estimated unbiasedly without estimating σ^2_i, $i \in s$. Thus if the sample of fsu's s is selected by a without replacement sampling scheme and the ssu's are selected by systematic sampling scheme, the variance of $\widehat{Y}_{ts}$ cannot be estimated unbiasedly because an unbiased estimator of σ^2_i based on the systematic sampling scheme is not available. On the other hand, if the fsu's are selected using either the PPSWR or SRSWR methods and the ssu's are selected by systematic sampling, then the variance of $\widehat{Y}_{ts}$ can be estimated unbiasedly because unbiased estimators of σ^2_i are not required for estimating the variance of $\widehat{Y}_{ts}$.

13.5 MODIFICATION OF VARIANCE ESTIMATORS

Srinath and Hidiroglou (1980), and Arnab (1988), proposed the following modifications of variance estimators when fsu's are selected by without replacement sampling procedure.

13.5.1 Srinath and Hidiroglou Modification

From each of the selected sample $s_i(i \in s)$, a subsample $\widetilde{s}_i$ of ssu's of suitable size $\widetilde{m}_i$ (to be determined) is selected independently by some suitable sampling scheme, and an unbiased estimator $\widehat{\widehat{Y}}_i$ of Y_i is obtained from $\widetilde{s}_i$, sacrificing observations within s_i but outside $\widetilde{s}_i$. Then,

$$V\left(\widehat{Y}_{ts}\right) = \sum_{i \in U} \alpha_i Y_i^2 + \sum_{i \neq} \sum_{j \in U} \alpha_{ij} Y_i Y_j + \sum_{i \in U} \beta_i \sigma_i^2$$

given in Eq. (13.3.2) can be estimated unbiasedly as a homogeneous quadratic estimator of $\widehat{\widehat{Y}}_i$ in the following form

$$\widetilde{V}\left(\widehat{Y}_{ts}\right) = \sum_{i \in s} c_{si} \widehat{\widehat{Y}}_i^2 + \sum_{i \neq} \sum_{j \in s} c_{sij} \widehat{\widehat{Y}}_i \widehat{\widehat{Y}}_j \tag{13.5.1}$$

provided, $\widetilde{m}_i$'s satisfy

$$(\alpha_i - 1)\widetilde{\sigma}_i^2 = \alpha_i \sigma_i^2 \quad \text{for } i = 1, \ldots, N \tag{13.5.2}$$

where $\widetilde{\sigma}_i^2 = V_2\left(\widehat{\widehat{Y}}_i\right)$ and c_{si} and c_{sij} meet the unbiasedness condition given in Eq. (13.3.3). For $b_{si} = 1/\pi_i$, we get $\alpha_i = 1/\pi_i$ and $\alpha_{ij} = \pi_{ij}/(\pi_i \pi_j) - 1$. Now for the choices of $c_{si} = (\alpha_i - 1)/\pi_i$ and $c_{sij} = (\alpha_{ij} - 1)/\pi_{ij}$, we illustrate how one obtains a $\widetilde{V}\left(\widehat{Y}_{ts}\right)$ in the following examples.

Example 13.5.1

Let s_i be an SRSWOR of size m_i, $\overline{y}(s_i)$ and $\overline{y}(\widetilde{s}_i)$ be the sample means of y_{ij} values based on the samples s_i and $\widetilde{s}_i$, respectively, $\widehat{Y}_i = M_i \overline{y}(s_i)$ and $\widehat{\widehat{Y}}_i = M_i \overline{y}(\widetilde{s}_i)$. Then $\sigma_i^2 = M_i^2(1/m_i - 1/M_i)S_{yi}^2$ and $\widetilde{\sigma}_i^2 = M_i^2(1/\widetilde{m}_i - 1/M_i)S_{yi}^2$. The condition (Eq. 13.5.2) yields

$$\widetilde{m}_i = m_i(1 - \pi_i)/(1 - m_i \pi_i / M_i) \tag{13.5.3}$$

Example 13.5.2

Let s_i of size m_i be chosen by PPSWR method with p_{ij} as normed size measures for the jth ssu of the ith fsu, and let $\widetilde{s}_i$ be an SRSWOR sub-sample of size $\widetilde{m}_i$. Let $\widehat{Y}_i = \sum_{j=1}^{m_i} y_{ij}/(m_i p_{ij})$ and $\widehat{\widehat{Y}}_i = \sum_{j=1}^{\widetilde{m}_i} y_{ij}/(\widetilde{m}_i p_{ij})$. Then, $\sigma_i^2 = \sum_{j=1}^{M_i} p_{ij}(y_{ij}/p_{ij} - Y_i)^2/m_i$ and $\widetilde{\sigma}_i^2 = m_i\sigma_i^2/\widetilde{m}_i$. The condition (Eq. 13.5.2) yields

$$\widetilde{m}_i = m_i(1-\pi_i) \tag{13.5.4}$$

13.5.2 Arnab Modification

In order to avoid loss of information, Arnab (1988) divided the sample s_i at random into r groups so that the kth group contains $m_i(k)$ ssu's, $\sum_{k=1}^{r} m_i(k) = m_i$. Let $\widehat{Y}_i(k)$ be an unbiased estimator of Y_i based on the kth group; then we can choose $\widehat{\widehat{Y}}_i = \sum_{k=1}^{r} w_i(k)\widehat{Y}_i(k)$ as a weighted average of $\widehat{Y}_i(k)$ with suitable weights $w_i(k)$'s $\left(0 \le w_i(k) \le 1, \sum_{k=1}^{r} w_i(k) = 1\right)$. The weights $w_i(k)$ and $m_i(k)$ are determined from the condition (Eq. 13.5.2).

Example 13.5.3

Let s_i be an SRSWOR with $\widehat{Y}_i = M_i\overline{y}(s_i)$ and $\widehat{Y}_i(k) = M_i\sum_{j=1}^{m_i(k)} y_{ij}/m_i(k)$. In this case

$$\sigma_i^2 = V(\widehat{Y}_i) = V\{M_i\overline{y}(s_i)\} = M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{yi}^2$$

and

$$\begin{aligned}
\widetilde{\sigma}_i^2 = V(\widehat{\widehat{Y}}_i) &= V\left(\sum_{k=1}^{r} w_i(k)\widehat{Y}_i(k)\right)\\
&= E\left[V\left(\sum_{k=1}^{r} w_i(k)\widehat{Y}_i(k)\middle| s_i\right)\right] + V\left[E\left(\sum_{k=1}^{r} w_i(k)\widehat{Y}_i(k)\middle| s_i\right)\right]\\
&= E\left(\sum_{k=1}^{r} w_i^2(k)V\{\widehat{Y}_i(k)|s_i\}\right.\\
&\quad \left. + \sum_{k\neq}^{r}\sum_{k'=1}^{r} w_i(k)w_i(k')Cov\{\widehat{Y}_i(k), \widehat{Y}_i(k')|s_i\}\right) + V(\widehat{Y}_i)
\end{aligned}$$

Now noting $V\left\{\widehat{Y}_i(k)\middle|s_i\right\} = M_i^2\left(\frac{1}{m_i(k)} - \frac{1}{m_i}\right)s_{yi}^2$ and $Cov\left\{\widehat{Y}_i(k), \widehat{Y}_i(k')\middle|s_i\right\} = -\frac{s_{yi}^2}{m_i}$ where $s_{yi}^2 = \frac{1}{m_i - 1}\sum_{j\in s_i}\{y_{ij} - \overline{y}(s_i)\}^2$, we find

$$\widetilde{\sigma}_i^2 = M_i^2 S_{yi}^2\left(\sum_{k=1}^{r}\frac{w_i^2(k)}{m_i(k)} - \frac{1}{M_i}\right).$$

The requirement (Eq. 13.5.2) yields

$$(\alpha_i - 1)\left(\sum_{k=1}^{r}\frac{w_i^2(k)}{m_i(k)} - \frac{1}{M_i}\right) = \alpha_i\left(\frac{1}{m_i} - \frac{1}{M_i}\right) \qquad (13.5.5)$$

The solution of $w_i(k)$'s in Eq. (13.5.5) may be obtained by a trial and error method. In particular, for $k = 2$, an easy solution is $w_i(1) = w_i(2) = 1/2$, $m_i(1) = m_i\left(1 + \sqrt{\delta_i}\right)/2$, and $m_i(2) = m_i\left(1 - \sqrt{\delta_i}\right)/2$ with $\delta_i = 1 - (\alpha_i - 1)/(\alpha_i - m_i/M_i)$.

13.6 MORE THAN TWO-STAGE SAMPLING

Here we will define operators E_L and V_L as overall expectation and variance later than the first stage of sampling, respectively. The selected sample of fsu's will be denoted by s as in earlier sections. If the ith fsu is included in the sample s, the sub-sample selected at the second and subsequent stages of sampling from the final (ultimate) stage from the ith fsu will be denoted by s_{iL}. Let $\widehat{Y}_{iL}$ be an unbiased estimator of Y_i based on s_{iL}, then an unbiased estimator of the total Y will be denoted as

$$\widehat{Y}_L = \sum_{i\in s} b_{si}\widehat{Y}_{iL} \qquad (13.6.1)$$

where b_{si}'s satisfy the unbiasedness condition $\sum_{s\supset i} b_{si}p(s) = 1$.

Now

$$\begin{aligned}
E\left(\widehat{Y}_L\right) &= E_1\left[\sum_{i\in s} b_{si}E_L\left(\widehat{Y}_{iL}\right)\right] \\
&= E_1\left(\sum_{i\in s} b_{si}Y_i\right) \\
&= Y
\end{aligned}$$

The expression for the variance of $\widehat{Y}_L$ and its unbiased estimator may be obtained by using Theorems 13.3.1 and 13.3.2 by writing σ_{iL}^2 in place of σ_i^2 and they are given as follows:

$$V\left(\widehat{Y}_L\right) = \sum_{i\in U} \alpha_i Y_i^2 + \sum_{i\neq}\sum_{j\in U} \alpha_{ij} Y_i Y_j + \sum_{i\in U} \beta_i \sigma_{iL}^2 \tag{13.6.2}$$

and

$$\widehat{V}\left(\widehat{Y}\right) = Q_s\left(\widehat{Y}_L\right) + \widehat{\Phi}_{sL} \tag{13.6.3}$$

where $Q_s\left(\widehat{Y}_L\right) = \sum_{i\in s} c_{si} \widehat{Y}_{iL}^2 + \sum_{i\neq}\sum_{j\in s} c_{sij} \widehat{Y}_{iL} \widehat{Y}_{jL}$ and $\widehat{\Phi}_{sL}$ are unbiased estimators of $Q\left(\widehat{Y}_L\right) = \sum_{i\in U} \alpha_i Y_i^2 + \sum_{i\neq}\sum_{j\in U} \alpha_{ij} Y_i Y_j$ and $\sum_{i\in U} \sigma_{iL}^2$, respectively; c_{si} and c_{sij} must satisfy $\sum_{s\supset i} c_{si} p(s) = \alpha_i$ and $\sum_{s\supset i,j} c_{sij} p(s) = \alpha_{ij}$ as in Theorem 13.3.2.

13.6.1 Three-Stage Sampling

Suppose a population is divided into N fsu's. The ith fsu consists of M_i ssu's and the jth ssu of the ith fsu consists of M_{ij} tsu's. The total number of tsu's is $M = \sum_{i=1}^{N}\sum_{j=1}^{M_i} M_{ij}$. Let a sample s of n fsu's be selected by SRSWOR method and if the ith fsu is selected in s, then a sub-sample s_i of size m_i ssu's is selected by SRSWOR method from the M_i ssu's of the ith fsu. Finally, if the jth ssu of the ith fsu is selected in the sample s_i, then a sub-sample s_{ij} of m_{ij} tsu's is selected by SRSWOR method from M_{ij} tsu's that belong to the ith fsu and jth ssu. Let y_{ijk} be the value of the study variable for the kth tsu of the jth ssu of ith fsu. Let $Y_{ij} = \sum_{k=1}^{M_{ij}} y_{ijk}$, $Y_i = \sum_{j=1}^{M_i} Y_i$, $Y = \sum_{i=1}^{N} Y_i$, and $\overline{y}(s_{ij}) = \sum_{k\in s_{ij}} y_{ijk}\big/ m_{ij}$. Then an unbiased estimator of the total Y is given by

$$\widehat{Y}_T = \frac{N}{n}\sum_{i\in s} \widehat{Y}_{iL} \tag{13.6.4}$$

where $\widehat{Y}_{iL} = \dfrac{M_i}{m_i}\sum_{j\in s_{ij}} \widehat{Y}_{ij}$ and $\widehat{Y}_{ij} = \dfrac{M_{ij}}{m_{ij}}\overline{y}(s_{ij})$.

Clearly $\widehat{Y}_{ij}$ is an unbiased estimator for Y_{ij}, and $\widehat{Y}_{iL}$ is an unbiased estimator of Y_i. Eqs. (13.6.2) and (13.6.3) yield the variance of $\widehat{Y}_{iL}$ and its unbiased estimator, respectively, as follows:

$$\begin{aligned} V\left(\widehat{Y}_L\right) &= V_1\left(\frac{N}{n}\sum_{i\in s} Y_i\right) + \frac{N^2}{n^2}\left(\sum_{i\in s}\sigma_{iL}^2\right) \\ &= N^2(1-f)\frac{S_b^2}{n} + \frac{N}{n}\sum_{i\in U}\sigma_{iL}^2 \end{aligned} \tag{13.6.5}$$

and

$$\widehat{V}\left(\widehat{Y}_L\right) = N^2(1-f)\frac{s_b^2}{n} + \frac{N}{n}\sum_{i\in s}\widehat{\sigma}_{iL}^2 \tag{13.6.6}$$

where $S_b^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$, $s_b^2 = \dfrac{1}{n-1}\sum_{i\in s}\left(\widehat{Y}_i - \widehat{\overline{Y}}_s\right)^2$,

$\overline{Y} = Y/N$, $\widehat{\overline{Y}}_s = \sum_{i\in s}\widehat{Y}_i/n$, and $f = n/N$.

The expressions of σ_{iL}^2 and $\widehat{\sigma}_{iL}^2$ are obtained by using Eqs. (13.3.10) and (13.3.11) as

$$\sigma_{iL}^2 = V\left(\widehat{Y}_{iL}\right) = M_i^2(1-f_i)\frac{S_{bi}^2}{m_i} + \frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{M_{ij}^2(1-f_{ij})S_{yij}^2}{m_{ij}}.$$

and

$$\widehat{\sigma}_{iL}^2 = M_i^2(1-f_i)\frac{s_{bi}^2}{m_i} + \frac{M_i}{m_i}\sum_{j\in s_i}\frac{M_{ij}^2(1-f_{ij})s_{yij}^2}{m_{ij}}$$

where $S_{bi}^2 = \sum_{j=1}^{M_i}\left(Y_{ij} - \overline{Y}_i\right)^2/(M_i - 1)$, $s_{bi}^2 = \sum_{j\in s_i}\left(\widehat{Y}_{ij} - \widehat{\overline{Y}}_{si}\right)^2\Big/(m_i - 1)$,

$f_i = m_i/M_i$, $f_{ij} = m_{ij}/M_{ij}$, $\overline{Y}_i = \sum_{j=1}^{M_i} Y_{ij}/M_i$, $\widehat{\overline{Y}}_{si} = \sum_{j\in s_i}\widehat{Y}_{ij}/m_i$,

$$S_{yij}^2 = \sum_{k=1}^{M_{ij}}\left(y_{ijk} - \overline{Y}_{ij}\right)^2\Big/(M_{ij} - 1),$$

$$s_{yij}^2 = \sum_{k=1}^{m_{ij}}\left\{y_{ijk} - \overline{y}(s_{ij})\right\}^2\Big/(m_{ij} - 1), \text{ and } \overline{y}(s_{ij}) = \sum_{k\in s_{ij}} y_{ijk}/m_{ij}.$$

13.7 ESTIMATION OF MEAN PER UNIT

The mean per unit is defined by $\overline{\overline{Y}} = Y/M_0$, where M_0 is the total number of ultimate units and Y is the population total based on M_0 ultimate units. For a two-stage sampling, $Y = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$ and $M_0 = \sum_{i=1}^{N} M_i$. In case M_0 is known, an unbiased estimate for the mean per unit is given by $\widehat{\overline{\overline{Y}}}_{ts} = \widehat{Y}_{ts}/M_0$. But in most cases, M_0 is unknown and an estimate of $\overline{\overline{Y}}$ is obtained as

$$\widehat{\overline{\overline{Y}}}_{ts}(r) = \widehat{Y}_{ts}\Big/\widehat{M}_0 \tag{13.7.1}$$

where $\widehat{Y}_{ts}$ and $\widehat{M}_0$ are an unbiased estimators of Y and M_0, respectively. For two-stage sampling,

$$\widehat{\overline{\overline{Y}}}_{ts}(r) = \sum_{i\in s} b_{si}\widehat{Y}_i \Big/ \sum_{i\in s} b_{si}M_i$$

Since $\widehat{\overline{\overline{Y}}}_{ts}(r)$ is a ratio estimator, it is biased for $\overline{\overline{Y}}$. The approximate expressions for the bias and mean-square error (MSE) are readily obtained from Theorem 8.2.2 and Eq. (8.2.16) and they are given as follows:

$$\text{Bias of } \widehat{\overline{\overline{Y}}}_{ts}(r) = B\left[\widehat{\overline{\overline{Y}}}_{ts}(r)\right] \cong -\frac{Cov\left[\widehat{Y}_{ts}, \widehat{M}_0\right]}{M_0} \tag{13.7.2}$$

$$\text{Mean square error of } \widehat{\overline{\overline{Y}}}_{ts}(r) = M\left[\widehat{\overline{\overline{Y}}}_{ts}(r)\right] \cong \frac{V\left(\sum_{i\in s} b_{si}\widehat{D}_i\right)}{M_0^2} \tag{13.7.3}$$

where $\widehat{D}_i = \widehat{Y}_i - \overline{\overline{Y}}M_i$

Now
$$\begin{aligned} V\left(\sum_{i\in s} b_{si}\widehat{D}_i\right) &= E\left(\sum_{i\in s} b_{si}^2 V\left(\widehat{D}_i|s_i\right)\right) + V\left(\sum_{i\in s} b_{si}D_i\right) \\ &= E\left(\sum_{i\in s} b_{si}^2\sigma_i^2(d)\right) + V\left(\sum_{i\in s} b_{si}D_i\right) \\ &= \sum_{i\in U}\alpha_i D_i^2 + \sum_{i\neq}\sum_{j\in U}\alpha_{ij}D_iD_j + \sum_{i\in U}\beta_i\sigma_i^2(d) \end{aligned}$$

where $D_i = Y_i - \overline{\overline{Y}}M_i$, $\sigma_i^2(d) = V\left(\widehat{D}_i|s_i\right)$; α_i, α_{ij} and β_i are as in Theorem 13.3.1.

Hence

$$M\left[\widehat{\overline{\overline{Y}}}_{ts}(r)\right] \cong \frac{1}{M_0^2}\left[\sum_{i\in U}\alpha_i D_i^2 + \sum_{i\neq}\sum_{j\in U}\alpha_{ij}D_iD_j + \sum_{i\in U}\beta_i\sigma_i^2(d)\right] \tag{13.7.4}$$

An approximate unbiased estimator of $M\left[\widehat{\overline{\overline{Y}}}_{ts}(r)\right]$ is

$$\widehat{M}\left[\widehat{\overline{\overline{Y}}}_{ts}(r)\right] \cong \frac{1}{\widehat{M}_0^2}\left[\sum_{i\in s}\alpha_i \widehat{\widehat{D}}_i^2\Big/\pi_i + \sum_{i\neq}\sum_{j\in s}\alpha_{ij}\widehat{\widehat{D}}_i\widehat{\widehat{D}}_j + \sum_{i\in s}\widehat{\sigma}_i^2(d)\big/\pi_i\right] \tag{13.7.5}$$

where $\widehat{\widehat{D}}_i = \widehat{Y}_i - \widehat{\overline{\overline{Y}}}_{ts}(r)M_i$, and $\widehat{\sigma}_i^2(d)$ is an approximate unbiased estimator of $\sigma_i^2(d)$.

13.7.1 Simple Random Sampling Without Replacement

If s and s_i's are selected by SRSWOR method, we have $b_{si} = N/n$, $\widehat{\overline{\overline{Y}}}_{ts}(wor) = \sum_{i \in s} M_i \bar{y}(s_i) \big/ \sum_{i \in s} M_i$, and the expressions (Eqs. 13.7.4 and 13.7.5) reduce to

$$M\left[\widehat{\overline{\overline{Y}}}_{ts}(wor)\right] \cong \frac{1}{\overline{M}^2}\left[(1-f)S_D^2/n + \frac{1}{nN}\sum_{i=1}^{N} M_i^2(1-f_i)S_{di}^2/m_i\right] \quad (13.7.6)$$

$$\widehat{M}\left[\widehat{\overline{\overline{Y}}}_{ts}(wor)\right] \cong \frac{1}{\widehat{\overline{M}}^2}\left[(1-f)s_{\hat{D}}^2/n + \frac{1}{nN}\sum_{i \in s} M_i^2(1-f_i)s_{\hat{d}i}^2\Big/ m_i\right] \quad (13.7.7)$$

where $\overline{M} = \dfrac{M_0}{N}$, $S_D^2 = \dfrac{\sum_{i=1}^{N}\left(D_i - \overline{D}\right)^2}{N-1}$, $\overline{D} = \dfrac{\sum_{i=1}^{N} D_i}{N}$,

$$S_{di}^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(D_{ij} - \overline{D}_i\right)^2, D_{ij} = y_{ij} - \overline{\overline{Y}} M_{ij}, \overline{D}_i = \frac{D_i}{M_i},$$

$$s_{\hat{D}}^2 = \frac{\sum_{i \in s}\left(\widehat{D}_i - \widehat{\overline{\overline{D}}}\right)^2}{n-1}, \widehat{D}_i = \widehat{Y}_i - \widehat{\overline{\overline{Y}}}_{ts}(wor)M_i,$$

$$\widehat{\overline{\overline{D}}} = \frac{1}{n}\sum_{i \in s}\widehat{D}_i,\ s_{\hat{d}i}^2 = \frac{1}{m_i - 1}\sum_{j \in s_i}\left(\widehat{D}_{ij} - \widehat{\overline{\overline{D}}}_i\right)^2, \widehat{D}_{ij} = y_{ij} - \widehat{\overline{\overline{Y}}}_{ts}(wor)M_{ij},$$

and $\widehat{\overline{\overline{D}}}_i = \frac{1}{m_i}\sum_{j \in s_i}\widehat{D}_{ij}$.

13.8 OPTIMUM ALLOCATION

In the two-stage sampling considered in the earlier section, we assumed that if the ith fsu is selected in the sample s, a sub-sample s_i of size m_i would be selected from it by some suitable sampling procedure. The size m_i was assumed to be pre-assigned. In this section we will determine the optimal values of m_i's and n that minimize: (i) the cost of the survey, keeping the precision of the estimator at a certain level and (ii) the variance of the estimator, keeping the cost fixed to a certain level. Here we consider the following simple cost function considered by Cochran (1977).

$$C = c_0 + c_1 n + c_2\sum_{i \in s} m_i + c_3\sum_{i \in s} M_i \quad (13.8.1)$$

where c_0 is the overhead cost of the survey, which is fixed for the survey, independent of n and m_i's, c_1 is the cost per fsu, which is assumed to be fixed, c_2 is the cost for surveying an ssu, and c_3 is the cost of listing per ssu. Since the cost C is a random variable that depends on the fsu's selected in s, we minimisze (i) the variance of an estimator, keeping the expected cost $\overline{C} = E(C) = c_0 + c_1 n + c_2 \sum_{i \in U} m_i \pi_i + c_3 \sum_{i \in U} M_i \pi_i$ to a certain level C^* (say) or (ii) the expected cost $\overline{C}$, keeping the variance of the estimator fixed to a certain level V^* (say). Obviously the solutions of (i) and (ii) exist if the variance of the estimator can be expressed as an explicit function of n and m_i's. Consider the case of two-stage sampling where both the fsu's and ssu' are selected by SRSWOR method. In this situation we obtain from Eq. (13.3.10)

$$V\left[\widehat{Y}_{ts}(wor)\right] = N^2(1-f)\frac{S_b^2}{n} + \frac{N}{n}\sum_{i \in U}\frac{M_i^2(1-f_i)S_{yi}^2}{m_i}$$

Now noting $\pi_i = n/N$, we get for the SRSWOR sampling

$$\overline{C} = c_0 + n(\overline{c}_1 + c_2\overline{m}) \tag{13.8.2}$$

where $\overline{c}_1 = c_1 + c_3 \sum_{i \in U} M_i/N$, $\overline{m} = \dfrac{1}{N}\sum_{i \in U} m_i$.

13.8.1 Fixed Expected Cost

Let us write

$$V\left[\widehat{Y}_{ts}(wor)\right] = \frac{1}{n}\left(A_0 + \sum_{i \in U}\frac{A_i}{m_i}\right) - B \tag{13.8.3}$$

with $A_0 = N^2 S_b^2 - N\sum_{i \in U} M_i S_{yi}^2$, $A_i = NM_i S_{yi}^2$, and $B = NS_b^2$.

Consider

$$\phi = V\left[\widehat{Y}_{ts}(wor)\right] - \lambda\left(\overline{C} - c_0 - \overline{c}_1 n - c_2 n\overline{m}\right) \tag{13.8.4}$$

with λ as an undetermined Lagrange's multiplier.

Now differentiating ϕ with respect to m_i and n and equating them to zero yield

$$m_i = \frac{\sqrt{NA_i}}{n\sqrt{\lambda\, c_2}} \tag{13.8.5}$$

and

$$\lambda n^2 = \frac{A_0 + \sum_{i \in U}\frac{A_i}{m_i}}{\overline{c}_1 + \frac{c_2}{N}\sum_{i \in U} m_i} \tag{13.8.6}$$

Eq. (13.8.5) implies

$$\lambda n^2 = \frac{N \sum_{i \in U} \frac{A_i}{m_i}}{c_2 \sum_{i \in U} m_i} \tag{13.8.7}$$

Eqs. (13.8.6) and (13.8.7) yield

$$\sum_{i \in U} \frac{A_i}{m_i} = \frac{c_2 A_0}{\bar{c}_1 N} \sum_{i \in U} m_i \tag{13.8.8}$$

and

$$\lambda = \frac{A_0}{n^2 \bar{c}_1} \tag{13.8.9}$$

On substituting λ in Eq. (13.8.5) the optimum value of m_i is obtained as

$$opt\, m_i = m_{i0} = \sqrt{\frac{N\bar{c}_1}{c_2 A_0} A_i} \tag{13.8.10}$$

Finally, using the relation $C^* = \overline{C} = c_0 + n\left(\bar{c}_1 + c_2 \frac{1}{N} \sum_{i \in U} m_{i0}\right)$ and substituting m_{i0} from Eq. (13.8.10), the optimum value of n comes out as

$$opt(n) = n_0 = \frac{C^* - c_0}{\bar{c}_1 + \left(\sum_{i \in U} \sqrt{A_i}\right)\sqrt{\frac{\bar{c}_1 c_2}{N A_0}}} \tag{13.8.11}$$

13.8.2 Fixed Variance

Here we find the optimum values of m and n, which minimize the expected cost $\overline{C} = c_0 + n\left(\bar{c}_1 + c_2 \frac{1}{N} \sum_{i \in U} m_i\right)$, keeping the variance to a certain level $\widehat{V}\left[\widehat{Y}_{ts}(wor)\right] = V^* = \frac{1}{n}\left(A_0 + \sum_{i \in U} \frac{A_i}{m_i}\right) - B$. Following Section 13.8.1, the optimum values of m and n are obtained, respectively, as follows:

$$opt(m_i) = m_{i0} = \sqrt{\frac{N\bar{c}_1}{c_2 A_0} A_i} \tag{13.8.12}$$

$$opt(n) = n_0 = \left(\frac{A_0}{V^* + B}\right)\left[1 + \left(\sum_{i\in U}\sqrt{A_i}\right)\sqrt{\frac{c_2}{N\bar{c}_1 A_0}}\right] \qquad (13.8.13)$$

13.9 SELF -WEIGHTING DESIGN

A linear homogeneous unbiased estimator of the population total Y based on a sample s is defined as

$$\widehat{Y} = \sum_{i\in s} b_{si} y_i$$

where the coefficient b_{si} is free from y_i but may depend on the unit i and the selected sample s. b_{si}, the coefficient of y_i, is called the multiplier or weight associated with the unit i. An estimator with a constant weight for every unit of the selected sample is termed as a "self-weighting estimator." Hence a self-weighting estimator with the constant weight w is of the form $\widehat{Y} = w\sum_{i\in s} y_i$. A sampling design that yields unbiased estimators for the population mean or total with a constant weight will be called a "self-weighting design." The estimator $\widehat{\overline{Y}} = \frac{1}{n}\sum_{i\in s} y_i$ of the population mean $\overline{Y} = \sum_{i=1}^{N} y_i/N$ based on an SRSWOR or SRSWR sampling provides equal weight $1/n$ for each of the units, and hence $\widehat{\overline{Y}}$ is a self-weighting estimator and SRSWR and SRSWOR are self-weighting designs. The Horvitz–Thompson estimator (1952) $\widehat{Y}_{ht} = \sum_{i\in s}\frac{y_i}{\pi_i}$ is not a self-weighting estimator unless all π_i's are equal. In large-scale surveys we estimate a large number of parameters based on various characteristics. Calculations become much simpler, quicker, and less erroneous if a self -weighing estimator is used. The self-weighting estimator may be obtained by choosing the appropriate sampling design. In multistage sampling design, the estimator $\widehat{Y}_{ts}(ht) = \sum_{i\in s}\widehat{Y}_i/\pi_i = \sum_{i\in s}\frac{M_i}{\pi_i}\frac{1}{m_i}\sum_{j\in s_i} y_{ij}$ given in Eq. (13.3.6) becomes self-weighting if $\pi_i \propto M_i$, and $m_i = m$ is a constant for every $i = 1,\ldots, N$. In stratified sampling (see Chapter 7) the estimator $\widehat{\overline{Y}}_{st} = \frac{1}{N}\sum_{i=1}^{N}\frac{N_i}{n_i}\sum_{j\in s_i} y_{ij}$ becomes a self-weighting estimator under the proportional allocation, where $n_i \propto N_i$.

Example 13.9.1

A sample of 10 orchards from 60 orchards in a certain village is selected by SRSWOR method. From each of the selected orchards, samples of apple plants are selected again by SRSWOR method. The following table gives the production of apples from each of the selected sampled plants.

Selected orchards	Number of plants	Production of apples (in kg) for the sampled plants					
1	50	50	75	125	75	100	120
2	25	35	80	40	70		
3	70	60	100	30	50	75	120
4	20	50	40	40	40		
5	30	50	60	65	70		
6	45	45	50	70	60		
7	50	60	90	80	75	40	80
8	30	55	75	70	70		
9	25	125	100	150	80		
10	25	70	80	75	70		

Estimate the average production of apples per plant when the average number of plants per orchard is (i) known to be 40 and (ii) unknown. Estimate the standard errors of the estimators used in "i" and "ii."

Solution

Here we first prepare the following table.

Selected orchards	# plants	Sampled plants	Sample mean yield	Sample variance	$\widehat{Y}_i$	$\widehat{\sigma}_i^2$	$\widehat{\widehat{D}}_i$
	M_i	m_i	$\overline{y}_i$	s_{yi}^2	$M_i\overline{y}_i$	$M_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)s_{yi}^2$	$\widehat{Y}_i - \widehat{\widehat{\overline{Y}}}_{st}(wor)M_i$
1	50	6	90.83	854.17	4,541.67	313,194.44	961.1486487
2	25	4	56.25	489.58	1,406.25	64,257.81	−384.009009
3	70	6	72.50	1097.50	5,075	819,466.67	62.2747748
4	20	4	42.50	25.00	850.00	200.000	−582.207207
5	30	4	61.25	72.92	1,837.50	14,218.75	−310.810811
6	45	4	56.25	122.92	2,531.25	56,695.31	−691.216216
7	50	6	70.83	324.17	3,541.67	118,861.11	−38.8513513
8	30	4	67.50	75.00	2,025.00	14,625.00	−123.310811
9	25	4	113.75	922.92	2,843.75	121,132.81	1053.490991
10	25	4	73.75	22.92	1,843.75	3,007.8125	53.490991
Total	370				26,495.83	1,527,459.72	

Estimated total production of apples in the village is

$$\widehat{Y}_{ts}(wor) = \frac{N}{n}\sum_{i\in s}\widehat{Y}_i = \frac{60}{10}\times 26,495.83 = 158,974.98 \text{ kg}$$

(i) Here the average number of plants per orchard is $\overline{M} = 40$. Hence the average production of apples per plant is $\widehat{\overline{\overline{Y}}}{}^{*}_{ts}(wor) = \widehat{Y}_{ts}(wor)\Big/\left(N\overline{M}\right) = 158{,}974.98/(60\times 40) = 66.24$ kg.

An unbiased estimate of variance of $\widehat{\overline{\overline{Y}}}$ is

$$\begin{aligned}
\widehat{V}\left[\widehat{\overline{\overline{Y}}}{}^{*}_{ts}(wor)\right] &= \frac{1}{\overline{M}^2}\left[s_b^2 + \frac{1}{Nn}\sum_{i\in s}\widehat{\sigma}_i^2\right]\\
&= \frac{1}{\overline{M}^2}\left[\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\sum_{i\in s}\left(\widehat{Y}_i - \widehat{\overline{Y}}\right)^2 + \frac{1}{Nn}\sum_{i\in s}\widehat{\sigma}_i^2\right]\\
&= \frac{1}{40^2}\left[\left(\frac{1}{10}-\frac{1}{60}\right)\times 1,865,941.165 + \frac{1}{10\times 60}\times 1,527,459.72\right]\\
&= 98.78
\end{aligned}$$

Hence an estimated standard error of $\widehat{\overline{\overline{Y}}}{}^{*}_{ts}(wor)$ is $se\left[\widehat{\overline{\overline{Y}}}{}^{*}_{ts}(wor)\right] = \sqrt{\widehat{V}\left[\widehat{\overline{\overline{Y}}}{}^{*}_{ts}(wor)\right]} = 9.94 \text{ kg}$

(ii) In case the average number of plants per orchard $\overline{M}$ is unknown, an estimate of average production of apples per plant is $\widehat{\overline{\overline{Y}}}_{ts}(wor) = \widehat{Y}_{ts}(wor)\Big/\left(N\widehat{\overline{M}}\right) = \sum_{i\in s}\widehat{Y}_i\Big/\sum_{i\in s}M_i = 26,495.83/370 = 71.61$ kg.

Following Eq. (13.7.7), an approximate estimate of the MSE of $\widehat{\overline{\overline{Y}}}_{ts}(wor)$ is given by

$$\widehat{M}\left[\widehat{\overline{\overline{Y}}}_{ts}(wor)\right] \cong \frac{1}{\widehat{\overline{M}}^2}\left[\left(\frac{1}{n}-\frac{1}{N}\right)s_{\hat{D}}^2+\frac{1}{nN}\sum_{i\in s}\widehat{\sigma}_i^2\right]$$

$$= \frac{1}{(370/10)^2}\left[\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\sum_{i\in s}\left(\widehat{\widehat{D}}_i-\widehat{\widehat{\overline{D}}}\right)^2+\frac{1}{nN}\sum_{i\in s}\widehat{\sigma}_i^2\right]$$

$$= \frac{1}{37^2}\left[\left(\frac{1}{10}-\frac{1}{60}\right)\times 346,435.086+\frac{1}{10\times 60}\times 1,527,459.72\right]$$

$$= 22.95$$

Estimated standard error of $\widehat{\overline{\overline{Y}}}_{ts}(wor)$ is $se\left[\widehat{\overline{\overline{Y}}}_{ts}(wor)\right] = \sqrt{\widehat{M}\left[\widehat{\overline{\overline{Y}}}_{ts}(wor)\right]}$ $= 4.79$ kg

Example 13.9.2

To estimate the total production of paddy in a certain district, a sample of 10 villages was selected from 50 villages in the district by PPSWR method using the area under cultivation of paddy as measure of size variable. From each of the selected villages, a sample of 20 fields was selected by systematic sampling procedure. The data are given below.

Sampled villages	Area under cultivation (in acre)	Total number of fields	Sample mean production per field (000 kg)
1	12.5	100	30
2	10.0	250	20
3	12.0	160	25
4	10.0	300	40
5	8.0	150	30
6	7.5	85	25
7	12.5	100	50
8	7.5	150	30
9	7.5	100	45
10	12.5	105	50

Estimate the total production of paddy and also estimate the standard error of the estimator assuming total area under cultivation in the district is 750 acres.

Solution

Here total number of villages in the district $N = 50$ and number of sampled villages $n = 10$.

The estimated total yield of the sampled villages and normed size measures associated with the villages are given as follows:

Sampled villages (r)	**Area under paddy** $X(r)$	$M(r)$	**Average yield** $\overline{y}(r)$	$p(r)$	$\widehat{Y}(r) = M(r)\,\overline{y}(r)$	$\widehat{Y}(r)/p(r)$
1	12.5	100	30	0.016667	3000	180,000
2	10	250	20	0.013333	5000	375,000
3	12	160	25	0.016000	4000	250,000
4	10	300	40	0.013333	12,000	900,000
5	8	150	30	0.010667	4500	421,875
6	7.5	85	25	0.010000	2125	212,500
7	12.5	100	50	0.016667	5000	300,000
8	7.5	150	30	0.010000	4500	450,000
9	7.5	100	45	0.010000	4500	450,000
10	12.5	105	50	0.016667	5250	315,000

$$\text{Estimated the total yield of paddy} = \widehat{Y}_{ts}(hh) = \frac{1}{n}\sum_{r=1}^{n}\widehat{Y}(r)/p(r) \times 1000 \text{ kg}$$

$$= 385,437.5 \times 1000 \text{ kg}$$

$$= 385,437,500 \text{ kg}$$

Estimated standard error of $\widehat{Y}_{ts}(hh)$ is

$$Se\left[\widehat{Y}_{ts}(hh)\right] = \sqrt{\frac{1}{n(n-1)}\sum_{r=1}^{n}\left(\widehat{Y}(r)/p(r) - \frac{1}{n}\sum_{r=1}^{n}\widehat{Y}(r)/p(r)\right)^2} \times 1000$$

$$= \sqrt{4191823351} \times 1000 \text{ kg} = 64,744,292.03 \text{ kg}$$

13.10 EXERCISES

13.10.1 Consider a two-stage sampling procedure where a sample s of n fsu's is selected by the PPSWR method using normed size measure p_i attached to the ith unit, $i = 1,\ldots, N$; $\sum_{i=1}^{N} p_i = 1$, and let

$n_i(s)$ be the number of times the ith fsu appeared in the sample$s$. The selected fsu's are sub-sampled in the following manner.

(a) A single sub-sample of size m_i is selected from the ith fsu, and let $T_1 = \frac{1}{n}\sum_{i\in s}\frac{n_i(s)\widehat{Y}_i}{p_i}$ be an estimator of the population total Y, where $\widehat{Y}_i = M_i\overline{y}_i$ and $\overline{y}_i$ is the sample mean based on the m_i ssu's selected from the ith fsu.

(b) $n_i(s)$-independent sub-samples, each of size m_i, are selected from the ith fsu and let $T_2 = \frac{1}{n}\sum_{r=1}^{n}\frac{\widehat{Y}(r)}{p(r)}$, where $\widehat{Y}(r) = M_j\overline{y}_j(r)$ if the rth draw produces the jth fsu, and $\overline{y}_j(r)$ is the sample mean of the ssu's selected from the jth fsu.

(c) $m_in_i(s)$ $(\leq M_i)$ ssu's are selected from the ith fsu and let $T_2 = \frac{1}{n}\sum_{i\in s}\frac{n_i(s)\widehat{Y}_i^*}{p_i}$, where $\widehat{Y}_i^* = M_i\overline{y}_i^*$ and $\overline{y}_i^*$ = sample mean based on $m_in_i(s)$ ssu's of the ith fsu.

Show that (i) $E(T_1) = E(T_2) = E(T_3) = Y$ and (ii) $V(T_1) \geq V(T_2) \geq V(T_3)$. Find unbiased estimators of $V(T_1)$, $V(T_2)$, and $V(T_3)$ (Rao, 1961).

13.10.2 Consider a two-stage sampling scheme where a sample of n fsu's is selected by the PPSWOR method using normed size measure p_i attached to the ith unit, $i = 1,\ldots, N$. Let s be an unordered sample obtained from the selected fsu's. If the ith fsu is included in s, a sub-sample s_i of size m_i ssu's is selected from the ith fsu by SRSWOR method. Let $T = \frac{\sum_{i\in s} M_i\overline{y}_ip(s|i)}{P(s)}$, where $\overline{y}_i$ is the sample mean of ssu's selected from the ith fsu, $p(s|i)$ = conditional probability of selection of s given that the ith unit is selected at the first draw, and $p(s)$ is the probability of selection of the unordered sample s. Show that T is an unbiased estimator of Y. Derive $V(T)$, the variance of T and an unbiased estimator of $V(T)$.

13.10.3 Let a sample s of n fsu 's be selected from a population by some suitable sampling schemes with π_i and π_{ij} as inclusion probabilities for the i th and ith and jth $(i \neq j)$ units, respectively. If ith fsu is included in s, a sub-sample s_i of size m_i ssu's is selected by systematic sampling scheme and let t_i be the sample mean.

(a) Show that $T = \sum_{i \in s} M_i t_i / \pi_i$ is an unbiased estimator of the population total Y. Derive the expression for $V(T)$. Explain why $V(T)$ cannot be estimated unbiasedly.

(b) Instead of selecting a single systematic sample s_i, two independent systematic samples $s_i(1)$ and $s_i(2)$ each of sizes $m_i/2$ (assumed to be integer) are selected by a systematic sampling scheme. Let $t_i(1)$ and $t_i(2)$ be the sample mean of ssu's based on samples $s_i(1)$ and $s_i(2)$, respectively. Show that

(i) $T = \sum_{i \in s} \frac{M_i \bar{t}_i}{\pi_i}$ with $\bar{t}_i = \frac{t_i(1) + t_i(2)}{2}$ is an unbiased estimator of Y.

(ii) $V(T) = \frac{1}{2} \sum_{i \neq}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 + \sum_{i=1}^{N} \frac{M_i^2 V(\bar{t}_i | s)}{\pi_i}$

and

(iii) $\widehat{V}(T) = \frac{1}{2} \sum_{i \neq} \sum_{j \in s} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{M_i t_i}{\pi_i} - \frac{M_j t_j}{\pi_j} \right)^2 + \frac{1}{4} \sum_{i \in s} \frac{M_i^2 \{ \bar{t}_i(1) - \bar{t}_i(2) \}^2}{\pi_i}$ is an unbiased estimator of $V(T)$.

13.10.4 From a population of N fsu's, a sample of n fsu's is selected by the Lahiri–Midzuno–Sen (1951, 1952, 1953) sampling scheme with p_i as normed size measure for the ith unit. If the ith fsu is selected in the sample, a sub-sample of m_i ssu's is selected by SRSWOR method. Suggest an unbiased estimator of the population total. Derive its variance and an unbiased estimator of the variance.

13.10.5 From a population of N enumeration areas, a sample of n enumerations areas are selected by SRSWOR method. If the ith enumeration area is selected in the sample, a sample s_i of m_i households is selected from the M_i households of the ith enumeration area by SRSWOR method. Let P_i and $\widehat{P}_i$ denote the proportion of households that possess a certain attribute (viz. owning a house) in the entire ith enumeration area and sampled households, respectively. Prove that

(i) $\widehat{P} = \frac{1}{n\overline{M}} \sum_{i \in s} M_i \widehat{P}_i$ is an unbiased estimator of P, the population proportion of houseldholds owning a house where

$\overline{M} = \sum_{i=1}^{N} M_i / N$ is known.

(ii) $$V\left(\widehat{P}\right)=\frac{1}{\overline{M}^2}\left[\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}\left(M_iP_i-\overline{M}P\right)^2+\frac{1}{nN}\sum_{i=1}^{N}M_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\frac{M_iP_i(1-P_i)}{M_i-1}\right]$$

and

$$\widehat{V}\left(\widehat{P}\right)=\frac{1}{\overline{M}^2}\left[\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\sum_{i\in s}\left(M_i\widehat{P}_i-\overline{M}\widehat{P}\right)^2+\frac{1}{nN}\sum_{i=1}^{N}M_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\frac{m_i\widehat{P}_i\left(1-\widehat{P}_i\right)}{m_i-1}\right]$$

13.10.6 Consider the sampling scheme in Exercise 13.10.5, where $\overline{M}$ is unknown. Show that the estimator $\widehat{P}^* = \sum_{i\in s} M_i\widehat{P}_i \Big/ \sum_{i\in s} M_i$ is a biased estimator of P. Derive approximate expressions for the bias and MSE of $\widehat{P}^*$.

13.10.7 From a population of N fsu's, a sample s of n ssu's is selected by SRSWOR method. All the ssu's of the selected fsu's are mixed up to from $M_0 = \sum_{i\in s} M_i$ ssu's. From the M_0 ssu's, a sample of m_0 ssu's are selected by SRSWOR method. Suggest an unbiased estimator of the population total. Derive its variance and unbiased estimator of this variance.

13.10.8 To estimate the average monthly rent of two-bedroom apartments in a certain city, the city was divided into 50 blocks. A sample of five blocks is selected by SRSWOR method, and from each of the blocks, five apartments were selected by SRSWOR method. The data are given below.

Sampled blocks	Number of two-bedroom apartments	Monthly rent of the sampled apartments (in $)
1	150	3500, 2800, 3500, 4500, 2800
2	150	5000, 3250, 3500, 3000, 3500
3	175	5000, 4500, 4000, 4500, 2800
4	200	4250, 1800, 3500, 4500, 5000
5	150	3500, 2500, 2800, 3500, 4000

Estimate the average rent of the apartment and its standard error when (i) the total number of apartments in the city is known as 875 and (ii) total number of apartment is unknown.

13.10.9 To estimate the total number of HIV/AIDS patients treated in a certain district, a stratified two-stage sampling design was used. Four blocks were selected from each of the strata by PPSWR method using number of people in the block as the measure of size. From each of the selected blocks, five clinics were selected by a systematic sampling procedure. The data obtained are as follows.

Stratum	Total population (in 000)	Sampled blocks	Population (in 000)	Total number of patients treated
1	150	1	5.20	75, 20, 15, 20, 10
		2	8.75	50, 30, 25, 20, 20
		3	12.80	75, 50, 25, 30, 10
		4	10.25	75, 20, 15, 20, 10
2	175	1	8.75	80, 20, 35, 15, 15
		2	12.80	25, 10, 10, 18, 12
		3	15.20	45, 40, 10, 20, 20
		4	16.25	25, 10, 15, 10, 10

Estimate the total number of patients treated and obtain an estimate of its standard error.

13.10.10 In a locality, a sample of six enumeration areas is selected from 50 enumeration areas by SRSWOR method. From each of the selected enumeration areas, a sample of five households is selected by SRSWOR method. The data on the number of HIV infected people in the selected households are given below.

Enumeration area	Total number of households	Sampled households					
1	40	Size	3	4	2	4	5
		#HIV infected	1	2	0	0	1
2	30	Size	4	4	2	1	1
		#HIV infected	0	0	2	0	0
3	20	Size	5	6	4	5	3
		#HIV infected	2	0	2	1	2
4	50	Size	4	5	3	2	1
		#HIV infected	3	2	0	1	0
5	80	Size	5	4	2	4	3
		#HIV infected	2	2	1	2	0
6	75	Size	3	3	4	2	1
		#HIV infected	1	2	2	1	0

Estimate the proportion of HIV infected persons in the locality and its standard error when the total number of people M in the locality (i) is known as 7500 and (ii) is unknown.