# CHAPTER 10

# Two-Phase Sampling

## 10.1 INTRODUCTION

It has been explained in Chapters 5, 8, and 9 how one can use auxiliary information at the stages of estimation of parameters, selection of samples, stratification, and their combinations to increase the precision of the estimators. To use auxiliary information for estimation of the population characteristics through ratio and regression methods, one needs to know only the population total of the auxiliary variable. For using the auxiliary variable as a measure of size for the selection of sample, one needs to know the values of the auxiliary variable at unit level. For the purpose of stratification using an auxiliary variable, one needs to know the values of the auxiliary variable for a large number of units so that after formation of the strata, each unit of the population can be assigned to one of the strata. In some situations the auxiliary information is not available in advance but can be made available at little extra cost. For example, in household income and expenditure surveys, one can easily collect information regarding the household size at the time of listing the households by putting little or no additional cost.

Two-phase sampling, also known as double sampling, was introduced by Neyman (1938). In two-phase sampling, samples are selected in phases. In the first phase, a relatively large sample $s'$ of size $n'$ is selected by some sampling design $p'$, and only information on the auxiliary variable is collected. During the second phase, a sample $s$ of size $n$ is selected either from $s'$, the sample selected in the first phase, or from the entire population by using a suitable sampling design $p$, and information regarding the study and auxiliary variable is collected. Evidently, two-phase sampling is useful if the auxiliary information $x$ is relatively easy and cheaper to collect than the study variable $y$ as well as if the strength of the relationship between the variables $x$ and $y$ is high. If the sample is selected in more than two phases, the resulting sampling design is called a multiphase sampling. In this chapter we will consider the use of the first-phase sample at the stages of estimation of the parameter, stratification, selection of sample, and their combinations in the second phase.

*Survey Sampling Theory and Applications*
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00010-8

## 10.2 TWO-PHASE SAMPLING FOR ESTIMATION

In the first phase, a sample $s'$ of size $n'$ be selected from a population with probability $p(s')$ using a sampling design $p$ and the information of the auxiliary variable $x$ is obtained only. Let

$$t(s', x) = \sum_{i \in s'} b_{s'i} x_i \tag{10.2.1}$$

be an unbiased estimator for the population total $X\left( = \sum_{i=1}^{N} x_i \right)$ of the auxiliary variable $x$ based on the first-phase sample $s'$ and $b_{s'i}$'s are known constants satisfying the unbiasedness condition

$$\sum_{s' \supset i} b_{s'i} p'(s') = 1 \quad \forall i = 1, \ldots, N \tag{10.2.2}$$

In the second phase, a subsample $s$ of size $n$ is selected from $s'$ and an unbiased estimator of the total $Y\left( = \sum_{i=1}^{N} y_i \right)$ is obtained as

$$t(s, y) = \sum_{i \in s} b_{si} y_i \tag{10.2.3}$$

Here the weights $b_{si}$'s are suitably chosen to satisfy

$$E\{t(s, y) | s'\} = \sum_{i \in s'} b_{s'i} y_i \tag{10.2.4}$$

The condition (Eq. 10.2.4) requires

$$\sum_{s \supset i} b_{si} p(s) = b_{s'i} \quad \forall i \in s' \tag{10.2.5}$$

### 10.2.1 Difference Method of Estimation

A difference estimator for the total $Y$ is given by

$$\widehat{Y}_{dt} = t(s, y) - \lambda \{t(s, x) - t(s', x)\} \tag{10.2.6}$$

where $t(s, x) = \sum_{i \in s} b_{si} x_i$ and $\lambda$ is a suitably chosen constant.

Clearly, $\widehat{Y}_{dt}$ is unbiased for $Y$ for any chosen value of $\lambda$ because $E\{t(s,y)\} = E[E\{t(s,y)|s'\}] = E\{t(s',y)\} = Y$ and $E\{t(s,x)\} = E\{t(s',x)\} = X$.

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

$$V\left(\widehat{Y}_{dt}\right) = V\left\{E\left(\widehat{Y}_{dt}|s'\right)\right\} + E\left\{V\left(\widehat{Y}_{dt}|s'\right)\right\}$$
$$= V\{t(s', y)\} + E[V\{t(s, d_\lambda)|s'\}] \tag{10.2.7}$$

where $t(s', y) = \sum_{i\in s'} b_{s'i}y_i$, $t(s, d_\lambda) = \sum_{i\in s} b_{si}d_{\lambda i}$, and $d_{\lambda i} = y_i - \lambda x_i$.

An unbiased estimator of $V\left(\widehat{Y}_{dt}\right)$ is given by

$$\widehat{V}\left(\widehat{Y}_{dt}\right) = \widehat{V}\{t(s', y)\} + \widehat{V}\{t(s, d_\lambda)|s'\} \tag{10.2.8}$$

where $\widehat{V}\{t(s', y)\}$ and $\widehat{V}\{t(s, d_\lambda)|s'\}$ are unbiased estimators of $V\{t(s', y)\}$ and $V\{t(s, d_\lambda)|s'\}$, respectively.

#### 10.2.1.1 Arbitrary Sampling Design

Let the initial sample $s'$ be selected by a varying probability sampling scheme with $\pi_i(>0)$ and $\pi_{ij}(>0)$ as the inclusion probabilities of the $i$th, and $i$th and $j$th units, respectively. The subsample $s$ is selected from $s'$ by simple random sampling without replacement (SRSWOR) method. In this case, if we choose, $t(s', x) = \sum_{i\in s'} \frac{x_i}{\pi_i}$ and $t(s, y) = \frac{n'}{n}\sum_{i\in s} \frac{y_i}{\pi_i}$, we get

$$\widehat{Y}_{dt} = \frac{n'}{n}\sum_{i\in s}\frac{y_i}{\pi_i} - \lambda\left(\frac{n'}{n}\sum_{i\in s}\frac{x_i}{\pi_i} - \sum_{i\in s'}\frac{x_i}{\pi_i}\right) \tag{10.2.9}$$

$$V\left\{E\left(\widehat{Y}_{dt}|s'\right)\right\} = V\{t(s', y)\}$$
$$= V\left(\sum_{i\in s'}\frac{y_i}{\pi_i}\right)$$
$$= \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{10.2.10}$$

and

$$E[V\{t(s, d_\lambda)|s'\}] = n'\frac{(n'-n)}{n(n'-1)}E\left[\sum_{i\in s'}\frac{d_{\lambda i}^2}{\pi_i^2} - n'\left(\frac{1}{n'}\sum_{i\in s'}\frac{d_{\lambda i}}{\pi_i}\right)^2\right]$$
$$= n'\frac{(n'-n)}{n(n'-1)}\left[\left(\sum_{i\in U}\frac{d_{\lambda i}^2}{\pi_i} - \frac{D_\lambda^2}{n'}\right) - \frac{1}{n'}\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{d_{\lambda i}}{\pi_i} - \frac{d_{\lambda j}}{\pi_j}\right)^2\right] \tag{10.2.11}$$

where $d_{\lambda i} = y_i - \lambda_i x_i$ and $D_\lambda = Y - \lambda X$.

Unbiased estimators of $V\{t(s', y)\}$ and $V(t(s, d_\lambda)|s')$ are, respectively, given by

$$\widehat{V}\{t(s', y)\} = \frac{n'}{n}\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{10.2.12}$$

and

$$\widehat{V}\{(t(s, d_\lambda)|s')\} = n'\frac{(n'-n)}{n(n-1)}\left[\sum_{i\in s}\frac{d_{\lambda i}^2}{\pi_i^2} - \frac{1}{n}\left(\sum_{i\in s}\frac{d_{\lambda i}}{\pi_i}\right)^2\right] \tag{10.2.13}$$

The expressions from Eq. (10.2.10) to Eq. (10.2.13) yield the variance of $\widehat{Y}_{dt}$ and its unbiased estimators as follows:

$$V\left(\widehat{Y}_{dt}\right) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + n'\frac{(n'-n)}{n(n'-1)} \times\left[\left(\sum_{i\in U}\frac{d_{\lambda i}^2}{\pi_i} - \frac{D_\lambda^2}{n'}\right) - \frac{1}{n'}\left\{\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{d_{\lambda i}}{\pi_i} - \frac{d_{\lambda j}}{\pi_j}\right)^2\right\}\right] \tag{10.2.14}$$

and

$$\widehat{V}\left(\widehat{Y}_{dt}\right) = \frac{n'}{n}\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + n'\frac{(n'-n)}{n(n-1)}\sum_{i\in s}\left(\frac{d_{\lambda i}}{\pi_i} - \frac{1}{n}\sum_{i\in s}\frac{d_{\lambda i}}{\pi_i}\right)^2 \tag{10.2.15}$$

#### *10.2.1.2 Simple Random Sampling Without Replacement*

In case, both the samples $s'$ and $s$ are selected by SRSWOR method, we get $\pi_i = n'/N$ and $\pi_{ij} = n'(n'-1)/\{N(N-1)\}$. In this situation the expressions (Eqs. 10.2.9, 10.2.14, and 10.2.15) reduce to

$$\widehat{Y}_{dt} = N\left[\overline{y}(s) - \lambda(\overline{x}(s) - \overline{x}(s'))\right] \tag{10.2.16}$$

$$V\left(\widehat{Y}_{dt}\right) = N^2\left[\left(\frac{N-n'}{Nn'}\right)S_y^2 + \frac{(n'-n)}{nn'}S_{d\lambda}^2\right] \tag{10.2.17}$$

and

$$V\left(\widehat{Y}_{dt}\right) = N^2\left[\left(\frac{N-n'}{Nn'}\right)s_y^2 + \frac{(n'-n)}{nn'}s_{d\lambda}^2\right] \tag{10.2.18}$$

where $\overline{x}(s') = \sum_{i \in s'} x_i/n'$, $\overline{x}(s) = \sum_{i \in s} x_i/n$, $\overline{y}(s) = \sum_{i \in s} y_i/n$,

$S_y^2 = \sum_{i \in U} \left(y_i - \overline{Y}\right)^2/(N-1)$, $S_{d\lambda}^2 = \sum_{i \in U} \left(d_{\lambda i} - \overline{D}_\lambda\right)^2/(N-1)$,

$s_y^2 = \sum_{i \in s} \left(y_i - \overline{y}_s\right)^2/(n-1)$, $s_{d\lambda}^2 = \sum_{i \in s} \left(d_{\lambda i} - \overline{d}_{\lambda s}\right)^2/(n-1)$,

$\overline{Y} = Y/N$, $\overline{D}_\lambda = D_\lambda/N$, and $\overline{d}_{\lambda s} = \sum_{i \in s} d_{\lambda i}/n$.

#### 10.2.1.3 Efficiency Under Simple Random Sampling Without Replacement

Let us consider the following simple cost function

$$C = c'n' + cn \tag{10.2.19}$$

where $C$ is the total cost of the survey excluding the fixed cost, $c'$ and $c$ denote, respectively, the cost for collecting information on $x$ and $y$ from a unit.

The cost $c$ is expected to be much higher than $c'$. If information regarding the study variable $y$ is collected only through direct survey (without using two-phase sampling) keeping $C$ as $C_0$, then the sample size for the direct survey would be

$$n_0 = C_0/c = (c'/c)n' + n \tag{10.2.20}$$

The variance of the sample mean $\overline{y}_{dr} = \sum_{i \in s_0} y_i/n_0$ based on a sample $s_0$ of size $n_0$ selected by SRSWOR method is given by

$$V\left(\overline{y}_{dr}\right) = \left(\frac{1}{n_0} - \frac{1}{N}\right) S_y^2$$

Two-phase sampling will be more precise than direct sampling if

$$V\left(\widehat{Y}_{dt}\right) < V\left(N\overline{y}_{dr}\right)$$

$$\text{i.e., if } N^2\left[\left(\frac{N-n'}{Nn'}\right)S_y^2 + \frac{(n'-n)}{nn'}\left(S_y^2 + \lambda^2 S_x^2 - 2\lambda\rho S_x S_y\right)\right] < N^2\left(\frac{1}{n_0} - \frac{1}{N}\right)S_y^2$$

$$\text{i.e., if } \rho > \frac{1}{2}\left[h + \left\{h\left(1 - \frac{n}{n'}\right)\left(1 + \frac{nc}{n'c'}\right)\right\}^{-1}\right] \tag{10.2.21}$$

where $\rho$ is the correlation coefficient between $x$ and $y$ and $h = \lambda S_x/S_y$.

#### 10.2.1.4 Probability Proportional to Size With Replacement Sampling

Let the initial sample $s'$ be selected by probability proportional to size with replacement (PPSWR) method using $z_i$ as a measure of size for the $i$th unit and the second phase sample $s$ be selected from $s'$ by SRSWOR method, treating all the units in $s'$ are distinct. In this case

$$\widehat{Y}_{dt} = \frac{1}{n}\sum_{i\in s}\frac{y_i}{q_i} - \lambda\left(\frac{1}{n}\sum_{i\in s}\frac{x_i}{q_i} - \frac{1}{n'}\sum_{i\in s'}\frac{y_i}{q_i}\right) \tag{10.2.22}$$

where $q_i = z_i/Z$ and $Z = \sum_{i\in U} z_i$ and $\sum_{i\in s}$ denotes the sum over the units in $s$ with repetition. Clearly, $\widehat{Y}_{dt}$ in Eq. (10.2.22) is unbiased for $Y$ and

$$\begin{aligned} V\left\{E\left(\widehat{Y}_{dt}|s'\right)\right\} &= V\left(\frac{1}{n'}\sum_{i\in s'}\frac{y_i}{q_i}\right) \\ &= \frac{1}{n'}\sum_{i\in U} q_i\left(\frac{y_i}{q_i} - Y\right)^2; \end{aligned} \tag{10.2.23}$$

$$\begin{aligned} E\left\{V\left(\widehat{Y}_{dt}|s'\right)\right\} &= \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{1}{n'-1}E\sum_{i\in s'}\left\{\frac{d_{\lambda i}}{q_i} - \frac{1}{n'}\left(\sum_{i\in s'}\frac{d_{\lambda i}}{q_i}\right)\right\}^2 \\ &= \left(\frac{1}{n} - \frac{1}{n'}\right)\sum_{i\in U} q_i\left(\frac{d_{\lambda i}}{q_i} - D_\lambda\right)^2 \end{aligned} \tag{10.2.24}$$

Eqs. (10.2.23) and (10.2.24) yield

$$V\left(\widehat{Y}_{dt}\right) = \frac{1}{n'}\sum_{i\in U} q_i\left(\frac{y_i}{q_i} - Y\right)^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)\sum_{i\in U} q_i\left(\frac{d_{\lambda i}}{q_i} - D_\lambda\right)^2 \tag{10.2.25}$$

Unbiased estimators of $V\left\{E\left(\widehat{Y}_{dt}|s'\right)\right\}$ and $E\left\{V\left(\widehat{Y}_{dt}|s'\right)\right\}$ are given, respectively, by

$$\widehat{V}\left\{E\left(\widehat{Y}_{dt}|s'\right)\right\} = \frac{1}{n'(n-1)}\sum_{i\in s}\left(\frac{y_i}{q_i} - \frac{1}{n}\sum_{i\in s}\frac{y_i}{q_i}\right)^2$$

and

$$\widehat{V}\left(\widehat{Y}_{dt}|s'\right) = \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{1}{n-1}\sum_{i\in s}\left(\frac{d_{\lambda i}}{q_i} - \frac{1}{n}\sum_{i\in s}\frac{d_{\lambda i}}{q_i}\right)^2$$

Hence an unbiased estimator for $V\left(\widehat{Y}_{dt}\right)$ is

$$\widehat{V}\left(\widehat{Y}_{dt}\right) = \frac{1}{n'(n-1)} \sum_{i \in s} \left( \frac{y_i}{q_i} - \frac{1}{n} \sum_{i \in s} \frac{y_i}{q_i} \right)^2 + \left( \frac{1}{n} - \frac{1}{n'} \right) \frac{1}{n-1} \sum_{i \in s} \left( \frac{d_{\lambda i}}{q_i} - \frac{1}{n} \sum_{i \in s} \frac{d_{\lambda i}}{q_i} \right)^2 \tag{10.2.26}$$

Under the cost function, (Eq. 10.2.19) two-phase sampling will be more precise than the single phase using the same cost if

$$\frac{1}{n'} \sum_{i \in U} q_i \left( \frac{y_i}{q_i} - Y \right)^2 + \left( \frac{1}{n} - \frac{1}{n'} \right) \sum_{i \in U} q_i \left( \frac{d_{\lambda i}}{q_i} - D_\lambda \right)^2 < \frac{1}{n_0} \sum_{i \in U} q_i \left( \frac{y_i}{q_i} - Y \right)^2$$

$$\text{i.e., if } \delta > \frac{1}{2} \left[ h^* + \left\{ h^* \left( 1 - \frac{n}{n'} \right) \left( 1 + \frac{nc}{n'c'} \right) \right\}^{-1} \right] \tag{10.2.27}$$

where

$$\delta = \frac{\sum_{i \in U} q_i \left( \frac{y_i}{q_i} - Y \right) \left( \frac{x_i}{q_i} - X \right)}{\sqrt{\sum_{i \in U} q_i \left( \frac{y_i}{q_i} - Y \right)^2} \sqrt{\sum_{i \in U} q_i \left( \frac{x_i}{q_i} - X \right)^2}}$$

and

$$h^* = \lambda \sqrt{\sum_{i \in U} q_i \left( \frac{x_i}{q_i} - X \right)^2} \Bigg/ \sqrt{\sum_{i \in U} q_i \left( \frac{y_i}{q_i} - Y \right)^2}$$

.

## 10.2.2 Ratio Method of Estimation

The ratio estimator for the population total in two-phase sampling is given by

$$\widehat{Y}_{Rt} = \frac{t(s, y)}{t(s, x)} t(s', x) \tag{10.2.28}$$

Like the ordinary ratio estimator $\widehat{Y}_R$ described in Eq. (8.3.1), $\widehat{Y}_{Rt}$ is a biased estimator for $Y$.

Now writing,

$$Cov\left\{\frac{t(s,y)}{t(s,x)}, t(s,x)|s'\right\} = E(t(s,y)|s') - \left[E\left\{\frac{t(s,y)}{t(s,x)}|s'\right\}\right]E(t(s,x)|s')$$
$$= E(t(s,y)|s') - \left[E\left\{\frac{t(s,y)}{t(s,x)}|s'\right\}\right]t(s',x) \tag{10.2.29}$$

and taking expectation both sides of Eq. (10.2.29) we get

$$E\left[Cov\left\{\frac{t(s,y)}{t(s,x)}, t(s,x)|s'\right\}\right] = Y - E\left[E\left\{\frac{t(s,y)}{t(s,x)}t(s',x)|s'\right\}\right] \tag{10.2.30}$$
$$= Y - E\left(\widehat{Y}_{Rt}\right)$$

Thus we have

Theorem 10.2.1
The bias of $\widehat{Y}_{Rt}$ under two-phase sampling is

$$B\left(\widehat{Y}_{Rt}\right) = -E\left[Cov\left\{\frac{t(s,y)}{t(s,x)}, t(s,x)|s'\right\}\right]$$

The above expression of bias is exact. However, it is not useful as it does not provide any elegant expression for any specific sampling design.

#### *10.2.2.1 Approximate Expression of Bias*

Let us define

$$e_x = \{t(s,x) - X\}/X, e_y = \{t(s,y) - Y\}/Y, \text{ and } e'_x = \{t(s',x) - X\}/X \tag{10.2.31}$$

Then,

$$\widehat{Y}_{Rt} = Y\frac{(1+e_y)}{(1+e_x)}\left(1+e'_x\right)$$
$$= Y(1+e_y)\left(1-e_x+e_x^2-\cdots\right)\left(1+e'_x\right) \tag{10.2.32}$$
$$= Y\left(1+e_y-e_x+e'_x+e_x^2-e_xe_y+e'_xe_y-e_xe'_x+\cdots\right)$$

The expression of bias of $\widehat{Y}_{Rt}$ is given by

$$\begin{aligned} B\left(\widehat{Y}_{Rt}\right) &= E\left(\widehat{Y}_{Rt} - Y\right) \\ &= YE\left(e_y - e_x + e'_x + e_x^2 - e_x e_y + e'_x e_y - e_x e'_x + \cdots\right) \end{aligned} \tag{10.2.33}$$

On neglecting the terms $E\left(e_x^i e_y^j e_x'^k\right)$ with $i+j+k>2$, assuming it small for large sample size $n$, an approximate expression of bias of $\widehat{Y}_{Rt}$ up to the first order of approximation comes out as

$$\begin{aligned} B\left(\widehat{Y}_{Rt}\right) &= E\left(\widehat{Y}_{Rt} - Y\right) \\ &= YE\left(e_y - e_x + e'_x + e_x^2 - e_x e_y + e'_x e_y - e_x e'_x\right) \\ &= Y\left[E\left(e_x^2 - e_x e_y\right) + E\left\{e'_x E(e_y|s')\right\} - e'_x E(e_x|s')\right] \\ &= Y\left[E\left(e_x^2 - e_x e_y\right) + E\left(e'_x e'_y - e_x'^2\right)\right] \\ &= Y\left[\frac{V\{t(s,x)\}}{X^2} - \frac{Cov\{t(s,x),t(s,y)\}}{XY} + \frac{Cov\{t(s',y),t(s',x)\}}{XY} - \frac{V\{t(s',x)\}}{X^2}\right] \\ &= Y\left[\frac{V\{t(s,x)\} - V\{t(s',x)\}}{X^2} - \frac{Cov\{t(s,x),t(s,y)\} - Cov\{t(s',y),t(s',x)\}}{XY}\right] \end{aligned} \tag{10.2.34}$$

Now substituting

$$\begin{aligned} V\{t(s,x)\} &= E[V\{t(s,x)|s'\}] + V[E\{t(s,x)|s'\}] \\ &= E[V\{t(s,x)|s'\}] + V\{t(s',x)\} \end{aligned}$$

and

$$\begin{aligned} Cov\{t(s,x),t(s,y)\} &= E[Cov\{t(s,x),t(s,y)|s'\}] + Cov[E\{t(s,x)|s'\},E\{t(s,y)|s'\}] \\ &= E[Cov\{t(s,x),t(s,y)|s'\}] + Cov\{t(s',x),t(s',y)\} \end{aligned}$$

in Eq. (10.2.34), the expression of $B\left(\widehat{Y}_{Rt}\right)$ reduces to

$$B\left(\widehat{Y}_{Rt}\right) = Y\left[\frac{E[V\{t(s,x)|s'\}]}{X^2} - \frac{E[Cov\{t(s,x),t(s,y)|s'\}]}{XY}\right] \tag{10.2.35}$$

#### 10.2.2.2 Approximate Expression of Mean Square Error

The mean square error of $\widehat{Y}_{Rt}$ is given by

$$\begin{aligned} M\left(\widehat{Y}_{Rt}\right) &= E\left(\widehat{Y}_{Rt} - Y\right)^2 \\ &= Y^2 E\left(e_y - e_x + e'_x + e_x^2 - e_x e_y + e'_x e_y - e_x e'_x + \cdots\right)^2 \end{aligned}$$

Neglecting the terms $E\left(e_x^i e_y^j e_x^k\right)$ with $i+j+k>2$, an approximate expression of $M\left(\widehat{Y}_{Rt}\right)$ up to the first order of approximation is obtained as

$$\begin{aligned} M\left(\widehat{Y}_{Rt}\right) &= Y^2 E\left(e_y^2 + e_x^2 + e_x'^2 - 2e_y e_x - 2e_x e_x'^2 + 2e_y e_x'^2\right) \\ &= Y^2 E\left[\left(e_y^2 + e_x^2 - 2e_y e_x\right) + \left\{e_x'^2 - 2E(e_x|s')e_x'^2 + 2E(e_y|s')e_x'^2\right\}\right] \\ &= Y^2 E\left[\left(e_y^2 + e_x^2 - 2e_y e_x\right) - \left(e_x'^2 - 2e'_y e'_x\right)\right] \\ &= \left[V\{t(s,y)\} + R^2 V\{t(s,x)\} - 2RCov\{t(s,y), t(s,x)\}\right] \\ &\quad - \left[R^2 V\{t(s',x)\} - 2RCov\{t(s',y), t(s',x)\}\right] \\ &= V\left[t(s,y) - R\ t(s,x)\right] - V\left[E\{t(s,y) - R\ t(s,x)|s'\right] + V\{t(s',y)\} \\ &= V\{t(s',y)\} + E\left[V\{t(s,y) - R\ t(s,x)|s'\}\right] \\ &= V\{t(s',y)\} + E\left[V\{t(s,d)|s'\}\right] \end{aligned} \tag{10.2.36}$$

where $t(s,d) = \sum_{i\in s} b_{si} d_i$ and $d_i = y_i - Rx_i$.

An approximate expression for an unbiased estimator of $M\left(\widehat{Y}_{Rt}\right)$ is given by

$$\widehat{M}\left(\widehat{Y}_{Rt}\right) = \widehat{V}\{t(s',y)\} + \widehat{V}\left\{t\left(s,\widehat{d}\right)|s'\right\} \tag{10.2.37}$$

where $\widehat{V}\{t(s',y)\}$ is an unbiased estimator of $V\{t(s',y)\}$ and $\widehat{V}\left\{t\left(s,\widehat{d}\right)|s'\right\}$ is an approximate unbiased estimator of $V\{t(s,d)|s'\}$, which is obtained by replacing $d_i$ with its estimate $\widehat{d}_i = y_i - \widehat{R}x_i$ in the expression of $\widehat{V}\{t(s,d)|s'\}$, an unbiased estimator of $V\{t(s,d)|s'\}$.

Eqs. (10.2.35)−(10.2.37) yield

Theorem 10.2.2

Approximate expressions for

(i) Bias of $\widehat{Y}_{Rt}$ is

$$B\left(\widehat{Y}_{Rt}\right) = Y\left[\frac{E[V\{t(s,x)|s'\}]}{X^2} - \frac{E[Cov\{t(s,x),t(s,y)|s'\}]}{XY}\right]$$

(ii) Mean square error of $\widehat{Y}_{Rt}$ is

$$M\left(\widehat{Y}_{Rt}\right) = V\{t(s',y)\} + E[V\{t(s,d)|s'\}]$$

(iii) Estimator of $M\left(\widehat{Y}_{Rt}\right)$ is

$$\widehat{M}\left(\widehat{Y}_{Rt}\right) = \widehat{V}\{t(s',y)\} + \widehat{V}\left\{t\left(s,\widehat{d}\right)|s'\right\}$$

### *10.2.2.3 Simple Random Sampling Without Replacement*

Consider the situation when both the samples $s$ and $s'$ are selected by SRSWOR method of sampling. In this case $t(s,x) = N\ \overline{x}(s)$, $t(s,y) = N\ \overline{y}(s)$, $t(s',x) = N\ \overline{x}(s')$, $t(s',y) = N\ \overline{y}(s')$, $t(s,d) = N\{\overline{y}(s) - R\,\overline{x}(s)\}$, $E[V\{N\ \overline{x}(s)|s'\}] = N^2\left(\frac{1}{n} - \frac{1}{n'}\right)S_x^2$, $V\{N\,\overline{y}(s')\} = N^2\left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2$, $E[Cov\{N\overline{x}(s), N\overline{y}(s)|s'\}] = N^2\left(\frac{1}{n} - \frac{1}{n'}\right)S_{xy}$ with $S_{xy} = \sum_{i\in U}(x_i - \overline{X})(y_i - \overline{Y})/(N-1)$, and $E[V(N\{\overline{y}(s) - R\,\overline{x}(s)\}|s')] = N^2\left(\frac{1}{n} - \frac{1}{n'}\right)\left(S_y^2 + R^2S_x^2 - 2RS_{xy}\right)$.

The expressions of $\widehat{Y}_{Rt}$, $B\left(\widehat{Y}_{Rt}\right)$, $\widehat{B}\left(\widehat{Y}_{Rt}\right)$, $M\left(\widehat{Y}_{Rt}\right)$, and $\widehat{M}\left(\widehat{Y}_{Rt}\right)$ appear as follows:

$$\widehat{Y}_{Rt} = N\frac{\overline{y}(s)}{\overline{x}(s)}\ \overline{x}(s')$$

$$B\left(\widehat{Y}_{Rt}\right) = N\left(\frac{1}{n} - \frac{1}{n'}\right)\overline{Y}\left(\frac{S_x^2}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}\,\overline{Y}}\right)$$

$$= N\left(\frac{1}{n} - \frac{1}{n'}\right)\overline{Y}\left(C_x^2 - \rho\ \ C_xC_y\right)$$

$$\widehat{B}\left(\widehat{Y}_{Rt}\right) = N\left(\frac{1}{n} - \frac{1}{n'}\right)\overline{y}(s)\left(\frac{s_x^2}{\{\overline{x}(s)\}^2} - \frac{s_{xy}}{\overline{x}(s)\overline{y}(s)}\right)$$

$$M\left(\widehat{Y}_{Rt}\right) = N^2\left[\left(\frac{1}{n'}-\frac{1}{N}\right)S_y^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)\left(S_y^2 + R^2S_x^2 - 2RS_{xy}\right)\right]$$

$$\widehat{M}\left(\widehat{Y}_{Rt}\right) = N^2\left[\left(\frac{1}{n'}-\frac{1}{N}\right)s_y^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)\left(s_y^2 + \widehat{R}^2s_x^2 - 2\widehat{R}s_{xy}\right)\right]$$

where $s_x^2 = \dfrac{\sum_{i\in s}\{x_i - \overline{x}(s)\}^2}{(n-1)}$, $s_y^2 = \dfrac{\sum_{i\in s}\{y_i - \overline{y}(s)\}^2}{n-1}$, $s_{xy} = \dfrac{\sum_{i\in s}\{x_i - \overline{x}(s)\}\{y_i - \overline{y}(s)\}}{n-1}$; $C_x$ and $C_y$ are the coefficient of variation of $x$ and $y$, respectively.

If the total cost of the survey is fixed, the two-phase sampling will be more precise than the single phase sampling under the cost function (Eq. 10.2.19) if

$$N^2\left[\left(\frac{1}{n'}-\frac{1}{N}\right)S_y^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)\left(S_y^2 + R^2S_x^2 - 2RS_{xy}\right)\right] < N^2\left(\frac{1}{n_0}-\frac{1}{N}\right)S_y^2$$

$$\text{i.e., if } \rho > \frac{1}{2}\left[\overline{h} + \left\{\overline{h}\left(1-\frac{n}{n'}\right)\left(1+\frac{nc}{n'c'}\right)\right\}^{-1}\right] \tag{10.2.38}$$

where $\overline{h} = C_x/C_y$.

#### 10.2.2.4 Optimal Allocation Under Simple Random Sampling Without Replacement

Here we will find the optimum values of $n$ and $n'$, which minimize $M\left(\widehat{Y}_{Rt}\right)$ when total cost of the survey $C = c'n' + cn$ is fixed to a certain level $C_0$. For this minimization problem, consider

$$\phi = M\left(\widehat{Y}_{Rt}\right) - \mu(C_0 - cn - c'n')$$

where $\mu$ is an undetermined Lagrange multiplier.

For large $N$, we can write

$$\phi = N^2\left[\frac{1}{n'}S_y^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)S_d^2\right] - \mu(C_0 - cn - c'n')$$

where $S_d^2 = S_y^2 + R^2S_x^2 - 2RS_{xy}$.

Now $\dfrac{\partial\phi}{\partial n} = 0$ and $\dfrac{\partial\phi}{\partial n'} = 0$ yield

$$n = NS_d/\sqrt{c\mu} \text{ and } n' = N\sqrt{S_y^2 - S_d^2}/\sqrt{c'\mu} \tag{10.2.39}$$

Substituting $n$ and $n'$ from Eq. (10.2.39) in the constraint $C_0 = c'n' + cn$, the optimum values of $n$ and $n'$ are derived as

$$\begin{aligned} n_{opt} &= \frac{C_0}{S_d\sqrt{c} + \sqrt{\left(S_y^2 - S_d^2\right)c'}} \frac{S_d}{\sqrt{c}} \quad \text{and} \\ n'_{opt} &= \frac{C_0}{S_d\sqrt{c} + \sqrt{\left(S_y^2 - S_d^2\right)c'}} \frac{\sqrt{S_y^2 - S_d^2}}{\sqrt{c'}} \end{aligned} \tag{10.2.40}$$

Placing the optimum values of $n$ and $n'$ in the expression of $M\left(\widehat{Y}_{Rt}\right)$ and taking $1/N \cong 0$, we arrive at the minimum value of $M\left(\widehat{Y}_{Rt}\right)$ for large $N$ as

$$M_{\min}\left(\widehat{Y}_{Rt}\right) = \frac{N^2\left[S_d\sqrt{c} + \sqrt{\left(S_y^2 - S_d^2\right)c'}\right]^2}{C_0} \tag{10.2.41}$$

## 10.2.3 Regression Method of Estimation

The regression estimator for the population total for two-phase sampling is

$$\widehat{Y}_{regt} = t(s,y) - B(s)[t(s,x) - t(s',x)] \tag{10.2.42}$$

where $B(s) = \dfrac{\widehat{Cov}\{t(s,y), t(s,x)|s'\}}{\widehat{V}\{t(s,x)|s'\}}$, $\widehat{Cov}\{t(s,y), t(s,x)|s'\}$ is an unbiased estimator of $Cov\{t(s,y),t(s,x) - t(s',x)\} = E[Cov\{t(s,y),t(s,x)|s'\}]$, and $\widehat{V}\{t(s,x)|s'\}$ is an unbiased estimator of $V\{t(s,x) - t(s',x)\} = E[V\{t(s,x)|s'\}]$.

### 10.2.3.1 Approximate Expressions of Bias and Mean Square Errors

Let $e_B = \dfrac{B(s) - B}{B}$ with $B = \dfrac{Cov\{t(s,y), t(s,x) - t(s',x)\}}{V\{t(s,x) - t(s',x)\}}$.

The bias of $\widehat{Y}_{regt}$ is

$$\begin{aligned} B\left(\widehat{Y}_{regt}\right) &= E\left(\widehat{Y}_{regt}\right) - Y \\ &= E\left[Y(1+e_y) - B(1+e_B)X\left(e_x - e'_x\right)\right] - Y \\ &= -B\ X\ E\left\{(e_B)\left(e_x - e'_x\right)\right\} \\ &= -E\{B(s), t(s,x) - t(s',x)\} \\ &= -E[Cov\{B(s), t(s,x)|s'\}] \end{aligned} \tag{10.2.43}$$

The mean square error of $\widehat{Y}_{regt}$ is given by

$$M\left(\widehat{Y}_{regt}\right) = E\left(\widehat{Y}_{regt} - Y\right)^2$$
$$= E\left[Y(1+e_y) - BX(1+e_B)\left(e_x - e'_x\right) - Y\right]^2$$

Now neglecting terms $E\left\{e_B^i e_x^j e_x'^k e_y^l\right\}$ for $i+j+k+l>2$, an approximate expression of $M\left(\widehat{Y}_{regt}\right)$ is obtained as

$$M\left(\widehat{Y}_{regt}\right) \cong V[t(s,y) - B\{t(s,x) - t(s',x)\}]$$
$$= V\{t(s',y)\} + E[V\{t(s,\in)|s'\}] \tag{10.2.44}$$

where $t(s,\in) = \sum_{i\in s} b_{si}E_i$ and $E_i = y_i - B\,x_i$.

An approximate unbiased estimator of $M\left(\widehat{Y}_{regt}\right)$ is

$$\widehat{M}\left(\widehat{Y}_{regt}\right) = \widehat{V}(t(s',y)) + \widehat{V}(t(s,\widehat{\in})|s') \tag{10.2.45}$$

where $\widehat{V}(t(s',y))$ is an unbiased estimator of $\widehat{V}(t(s,y))$ and $\widehat{V}(t(s,\widehat{\in})|s')$ is an approximate unbiased estimator of $V(t(s,\in)|s')$. The estimator $\widehat{V}(t(s,\widehat{\in})|s')$ is obtained by replacing $E_i$ by its estimate $\widehat{E}_i = y_i - B(s)x_i$ in the expression of $\widehat{V}(t(s,\in)|s')$, an unbiased estimator of $V(t(s,\in)|s')$.

#### 10.2.3.2 Arbitrary Sampling Design

Let the initial sample $s'$ and $s$ be selected as in Section 10.2.1.1 and

$$t(s',y) = \sum_{i\in s'}\frac{y_i}{\pi_i}, t(s',x) = \sum_{i\in s'}\frac{x_i}{\pi_i} \text{ and } t(s,y) = \frac{n'}{n}\sum_{i\in s}\frac{y_i}{\pi_i}.$$

In this case

$$\widehat{Y}_{regt} = \frac{n'}{n}\sum_{i\in s}\frac{y_i}{\pi_i} - B(s)\left(\frac{n'}{n}\sum_{i\in s}\frac{x_i}{\pi_i} - \sum_{i\in s'}\frac{x_i}{\pi_i}\right) \tag{10.2.46}$$

where

$$B(s) = \frac{\widehat{Cov}\left(\sum_{i\in s}\frac{y_i}{\pi_i}, \sum_{i\in s}\frac{x_i}{\pi_i}|s'\right)}{\widehat{V}\left(\sum_{i\in s}\frac{y_i}{\pi_i}|s'\right)} = \frac{\sum_{i\in s}\left(\frac{x_i}{\pi_i} - \frac{1}{n}\sum_{i\in s}\frac{x_i}{\pi_i}\right)\sum_{i\in s}\left(\frac{y_i}{\pi_i} - \frac{1}{n}\sum_{i\in s}\frac{y_i}{\pi_i}\right)}{\sum_{i\in s}\left(\frac{x_i}{\pi_i} - \frac{1}{n}\sum_{i\in s}\frac{x_i}{\pi_i}\right)^2}.$$

The expression of bias of $\widehat{Y}_{regt}$ is very complicated and hence it is omitted here. On the other hand, the expression of $M\left(\widehat{Y}_{regt}\right)$ is quite elegant and is given as follows:

$$
\begin{aligned}
M\left(\widehat{Y}_{regt}\right) &\cong V\left(\sum_{i\in s'}\frac{y_i}{\pi_i}\right)+E\left\{V\left(\frac{n'}{n}\sum_{i\in s}\frac{E_i}{\pi_i}|s'\right)\right\}\\
&=\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\\
&\quad+\frac{n'(n'-n)}{n}\frac{1}{n'-1}E\left[\sum_{i\in s'}\left\{\frac{E_i}{\pi_i}-\frac{1}{n'}\sum_{i\in s'}\frac{E_i}{\pi_i}\right\}^2\right]\\
&=\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2+\frac{n'(n'-n)}{n}\frac{1}{n'-1}\\
&\quad\times\left[E\left(\sum_{i\in s'}\frac{E_i^2}{\pi_i^2}\right)-\left\{V\left(\sum_{i\in s'}\frac{E_i}{\pi_i}\right)^2+\left(\sum_{i\in U}E_i\right)^2\right\}\Big/n'\right]\\
&=\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2+n'\frac{(n'-n)}{n(n'-1)}\\
&\quad\times\left[\left\{\sum_{i\in U}\frac{E_i^2}{\pi_i}-\left(\sum_{i\in U}E_i\right)^2\Big/n\right\}-\frac{1}{n'}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{E_i}{\pi_i}-\frac{E_j}{\pi_j}\right)^2\right]
\end{aligned}
\tag{10.2.47}
$$

where $E_i=y_i-Bx_i$ and $B=\dfrac{E\left\{Cov\left(\sum_{i\in s}\frac{y_i}{\pi_i},\sum_{i\in s}\frac{x_i}{\pi_i}\right)|s'\right\}}{E\left\{V\left(\sum_{i\in s}\frac{x_i}{\pi_i}\right)|s'\right\}}$.

An approximate unbiased estimator of $M\left(\widehat{Y}_{regt}\right)$ is given by

$$
\begin{aligned}
\widehat{M}\left(\widehat{Y}_{regt}\right)&\cong\frac{n'(n'-1)}{n(n-1)}\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\\
&\quad+n'\frac{(n'-n)}{n(n-1)}\left\{\sum_{i\in s}\frac{\widehat{E}_i^2}{\pi_i^2}-\frac{1}{n}\left(\sum_{i\in s}\frac{\widehat{E}_i}{\pi_i}\right)^2\right\}
\end{aligned}
\tag{10.2.48}
$$

with $\widehat{E}_i=y_i-B(s)\,x_i$.

#### 10.2.3.3 Simple Random Sampling Without Replacement

In case samples $s'$ and $s$ are selected by SRSWOR method we have

$$t(s',y)=N\frac{1}{n'}\sum_{i\in s'}y_i=N\,\overline{y}(s'),\ t(s',x)=N\frac{1}{n'}\sum_{i\in s'}x_i=N\overline{x}(s'),$$

$$t(s,y)=n'\frac{1}{n}\sum_{i\in s}y_i=n'\ \overline{y}(s),$$

$$B(s)\ =\frac{\widehat{C}ov\{\overline{y}(s),\overline{x}(s)|s'\}}{\widehat{V}\{\overline{x}(s)|s'\}}=\frac{s_{xy}}{s_x^2}\ \text{ and}$$

$$\widehat{Y}_{regt}=N\left[\overline{y}(s)-\frac{s_{xy}}{s_x^2}(\overline{x}(s)-\overline{x}(s'))\right] \tag{10.2.49}$$

An approximate bias of $\widehat{Y}_{regt}$ is obtained from Eq. (10.2.43) as

$$B\left(\widehat{Y}_{regt}\right)=-NE\left[Cov\left\{\frac{s_{xy}}{s_x^2},(\overline{x}(s)-\overline{x}(s'))|s\right\}\right]$$

$$=-NE\left[\frac{s_{xy}}{s_x^2}(\overline{x}(s)-\overline{x}(s'))\right]$$

Now writing $\varepsilon_{xy}=\dfrac{s_{xy}-S_{xy}}{S_{xy}},\varepsilon_{xx}=\dfrac{s_x^2-S_x^2}{S_x^2}$, and $B=\beta=\dfrac{S_{xy}}{S_x^2}$, we get

$$\begin{aligned}B\left(\widehat{Y}_{regt}\right)&=-N\beta E\left[(1+\varepsilon_{xy})(1+\varepsilon_{xx})^{-1}(\overline{x}(s)-\overline{x}(s'))\right]\\&\cong-N\beta E\left[(\varepsilon_{xy}-\varepsilon_{xx})(\overline{x}(s)-\overline{x}(s'))\right]\\&=-N\beta\left[\frac{E\left[Cov\left\{s_{xy},\overline{x}(s)\right\}|s'\right]}{S_{xy}}-\frac{E\left[\left\{s_x^2,\overline{x}(s)\right\}|s'\right]}{S_x^2}\right]\end{aligned} \tag{10.2.50}$$

Now using Sukhatme et al. (1984), we find

$$\begin{aligned}E\left[Cov\left\{s_{xy},\overline{x}(s)\right\}|s'\right]&=\frac{n'(n'-n)}{(n'-1)(n'-2)n}E\left[\frac{1}{n'}\sum_{i\in s'}\{x_i-\overline{x}(s')\}^2\{y_i-\overline{y}(s')\}\right]\\&=\frac{n'(n'-n)}{(n'-1)(n'-2)n}\mu_{21}\end{aligned} \tag{10.2.51}$$

and

$$E\left[\left\{s_x^2, \overline{x}(s)\right\}|s'\right] = \frac{n'(n'-n)}{(n'-1)(n'-2)n} E\left[\frac{1}{n'}\sum_{i\in s'}\left\{x_i - \overline{x}(s')\right\}^3\right] = \frac{n'(n'-n)}{(n'-1)(n'-2)n}\mu_{30} \tag{10.2.52}$$

where $\mu_{ij} = \sum_{i\in U}\left(x_i - \overline{X}\right)^i\left(y_i - \overline{Y}\right)^j \Big/ N$

Now substituting Eqs. (10.2.51) and (10.2.52) in Eq. (10.2.50), we find

$$B\left(\widehat{Y}_{regt}\right) \cong -\beta\frac{Nn'(n'-n)}{(n'-1)(n'-2)n}\left(\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^2}\right)$$

If $n'$ is large compared to $n$, the bias $B\left(\widehat{Y}_{regt}\right)$ can be approximated as

$$B\left(\widehat{Y}_{regt}\right) \cong -N\beta\left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^2}\right) \tag{10.2.53}$$

Following Eqs. (10.2.44) and (10.2.45), approximate expressions for the mean square error of $\widehat{Y}_{regt}$ and its unbiased estimator are obtained as

$$M\left(\widehat{Y}_{regt}\right) \cong N^2\left[\left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_E^2\right] \cong N^2\left[\left(\frac{1}{n'} - \frac{1}{N}\right) + \left(\frac{1}{n} - \frac{1}{n'}\right)(1-\rho^2)\right]S_y^2 \tag{10.2.54}$$

and

$$\widehat{M}\left(\widehat{Y}_{regt}\right) \cong N^2\left[\left(\frac{1}{n'} - \frac{1}{N}\right)s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)s_{\widehat{E}}^2\right] = N^2\left[\left(\frac{1}{n'} - \frac{1}{N}\right) + \left(\frac{1}{n} - \frac{1}{n'}\right)(1-r^2)\right]s_y^2 \tag{10.2.55}$$

where $S_E^2 = \sum_{i\in U}\left(E_i - \overline{E}\right)^2/(N-1)$, $s_{\widehat{E}}^2 = \sum_{i\in s}\left(\widehat{E}_i - \overline{\widehat{E}}\right)^2/(n-1)$, $\overline{E} = \sum_{i\in U}E_i/N$, $\overline{\widehat{E}} = \sum_{i\in s}\widehat{E}_i/n$, $\widehat{E}_i = y_i - \widehat{\beta}x_i$, and $r = s_{xy}/(s_x s_y)$ = sample correlation coefficient.

#### 10.2.3.4 Optimum Allocation

For large $N$, the expression (Eq. 10.2.54) reduces to

$$M\left(\widehat{Y}_{regt}\right) = N^2\left[\frac{(1-\rho^2)}{n} + \frac{\rho^2}{n'}\right]S_y^2 \tag{10.2.56}$$

The optimum values of $n$ and $n'$ that minimize (Eq. 10.2.56) for a given cost $C = C_0 = cn + c'n'$ are obtained by minimizing $N^2\left[\frac{(1-\rho^2)}{n} + \frac{\rho^2}{n'}\right]S_y^2$, subject to $C_0 = cn + c'n'$. For minimization consider the function

$$\phi = N^2\left[\frac{(1-\rho^2)}{n} + \frac{\rho^2}{n'}\right]S_y^2 - \mu(C_0 - cn - c'n')$$

where $\mu$ is an undetermined Lagrange multiplier.

$\frac{\partial\phi}{\partial n} = 0$ and $\frac{\partial\phi}{\partial n'} = 0$ yield the optimum values of $n$ and $n'$ are as follows:

$$n_{opt} = \frac{C_0}{\left(\sqrt{(1-\rho^2)}\sqrt{c} + \rho\sqrt{c'}\right)}\frac{\sqrt{1-\rho^2}}{\sqrt{c}}$$

and

$$n'_{opt} = \frac{C_0}{\left(\sqrt{(1-\rho^2)}\sqrt{c} + \rho\sqrt{c'}\right)}\frac{\rho}{\sqrt{c'}}$$

The value of $M\left(\widehat{Y}_{regt}\right)$ with the optimum values of $n$ and $n'$ is given by

$$M_{\min} = N^2\frac{\left(\sqrt{(1-\rho^2)}\sqrt{c} + \rho\sqrt{c'}\right)^2}{C_0}S_y^2 \tag{10.2.57}$$

For the fixed cost $C_0$, the value of $n_0 = C_0/c$. Hence, two-phase sampling with optimum allocation will be more precise than the single-phase sampling with the same cost for a large $N$

$$\begin{aligned}
&\text{if } M_{\min} < N^2\frac{1}{n_0}S_y^2\\
&\text{i.e. if } \left(\sqrt{(1-\rho^2)}\sqrt{c} + \rho\sqrt{c'}\right)^2 < c\\
&\text{i.e. if } 2\sqrt{1-\rho^2}\sqrt{cc'} < \rho^2(c-c')\\
&\text{i.e. if } \rho^2 > \frac{4cc'}{(c+c')^2}
\end{aligned} \tag{10.2.58}$$

Example 10.2.1

A sample of 50 plants was selected from a garden of 500 plants by SRSWOR method, and eye estimates of the heights of the selected plants were noted. From the selected sample of 50 plants, a subsample of 10 plants was selected by SRSWOR method and the heights of the selected plants were measured accurately. The data are presented as follows. Estimate the average height of the plants by using (i) ratio and (ii) regression method of estimation using the information that the sample mean height of eye estimation of the 50 plants was 10.5 m. Compare standard errors of the estimators used. (iii) Given that the cost of eye estimation and actual estimation of measuring height are \$10 and \$50, respectively, estimate the optimum proportions of actual measurement that should be made for the ratio and regression estimators.

Eye estimates and actual heights of plants.

| Plants | Eye estimate of height (in m) (x) | Actual height (in m) (y) |
|---|---|---|
| 1 | 15 | 13.85 |
| 2 | 10 | 9.75 |
| 3 | 8 | 10 |
| 4 | 8.5 | 10.5 |
| 5 | 12 | 14 |
| 6 | 6 | 7 |
| 7 | 6.5 | 8 |
| 8 | 15 | 14 |
| 9 | 10 | 12 |
| 10 | 8 | 7.5 |

Here $N = 500$, $n' = 50$, $n = 10$, $\overline{x}(s') = 10.5$, $\overline{x}(s) = 9.9$, $\overline{y}(s) = 10.66$, $s_x^2 = 10.267$, $s_y^2 = 7.337$, $s_{xy} = 7.907$, $\widehat{R} = 10.66/9.9 = 1.077$, $s_d^2 = 2.213$, and sample correlation coefficient $= r = 0.911$.

(i) Ratio estimation:

Estimated mean height of the plants under two-phase sampling scheme is

$$\widehat{\overline{Y}}_{Rt} = \frac{\overline{y}(s)}{\overline{x}(s)}\,\overline{x}(s') = (10.66/9.9) \times 10.5 = 11.306 \text{ m}$$

$$\text{Estimated MSE of } \widehat{\overline{Y}}_R = \widehat{M}\left(\widehat{\overline{Y}}_{Rt}\right) = \left(\frac{1}{n'} - \frac{1}{N}\right)s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)s_d^2$$

$$= \left[\left(\frac{1}{50} - \frac{1}{500}\right) \times 7.336 + \left(\frac{1}{10} - \frac{1}{50}\right) \times 2.213\right]$$

$$= 0.3091$$

$$\text{Estimated standard error of } \widehat{\overline{Y}}_{Rt} = SE\left(\widehat{\overline{Y}}_{Rt}\right) = \sqrt{\widehat{M}\left(\widehat{\overline{Y}}_{Rt}\right)} = 0.556 \text{ m.}$$

(ii) Regression Estimation:

Estimated mean height of plants under two-phase sampling is

$$\widehat{\overline{Y}}_{regt} = \overline{y}(s) - \frac{s_{xy}}{s_x^2}\{\overline{x}(s) - \overline{x}(s')\}$$

$$= 10.66 - (7.116/10.266) \times (9.9 - 10.5) = 11.122 \text{ m}$$

$$\text{Estimated MSE of } \widehat{\overline{Y}}_{regt} = \widehat{M}\left(\widehat{\overline{Y}}_{regt}\right) = \left[\left(\frac{1}{n'} - \frac{1}{N}\right) + \left(\frac{1}{n} - \frac{1}{n'}\right)(1 - r^2)\right]s_y^2$$

$$= \left[\left(\frac{1}{50} - \frac{1}{500}\right) + \left(\frac{1}{10} - \frac{1}{50}\right)(1 - 0.911^2)\right] \times 7.336 = 0.232$$

$$\text{Estimated standard error of } \widehat{\overline{Y}}_{regt} = SE\left(\widehat{\overline{Y}}_{regt}\right) = \sqrt{\widehat{M}\left(\widehat{\overline{Y}}_{regt}\right)}$$

$= 0.482$ m.

Efficiency of the regression estimator compared to the ratio estimator is

$$\frac{M\left(\widehat{\overline{Y}}_{Rt}\right)}{M\left(\widehat{\overline{Y}}_{regt}\right)} \times 100 = 133.307\%.$$

(iii) Optimum proportion of second-phase sample:

The optimum proportion for the ratio estimator is

$$n_{opt}\Big/n'_{opt} = \sqrt{\frac{c'}{c}\frac{s_d^2}{s_y^2 - s_d^2}} = \sqrt{\frac{c'}{c}\frac{s_d^2}{s_y^2 - s_d^2}} = \sqrt{\frac{10}{50} \times \frac{2.213}{5.214}} = 0.291$$

The optimum proportion for the regression estimator is

$$n_{opt}\Big/n'_{opt} = \sqrt{\frac{c'}{c}\frac{(1 - r^2)}{r^2}} = \sqrt{\frac{10}{50} \times \frac{(1 - 0.911^2)}{0.911^2}} = 0.202$$

## 10.3 TWO-PHASE SAMPLING FOR STRATIFICATION

It is well known that stratification is an important tool to increase precision of estimators. In stratification, the entire population is divided into a number of disjoint and exhaustive strata so that the each stratum becomes homogeneous with respect to the character under study. If no criterion for formation of strata is known in advance, we may use two-phase sampling for forming the strata. In the first phase, a relatively large sample $s'$ of size $n'$ is selected from the entire population by SRSWOR method, and information on the auxiliary variable (which can be used for stratification) is obtained. Here it is supposed that by noting the value of the auxiliary variable, one can classify the sample $s'$ into a predetermined number of "$H$" strata. Let $s'_h$ be the set of units of size $n'_h$ falling in the $h$th stratum. Here we make the assumption that the sample size $n'$ is so large that $n'_h$'s always exceeds 1 for every $h = 1,\ldots, H$; in other words, $Prob(n'_h \geq 1) = 1$. From each of the selected samples $s'_h$'s, subsamples $s_h$'s of sizes $n_h = \gamma_h n'_h$ are selected independently by SRSWOR method. Here $\gamma_h$'s are predetermined fractions and $n'_h$'s are assumed to be integers. Hence, for proportional allocation with fixed $n$, we get $\gamma_h = \gamma = n/n'$ and $n_h = nn'_h/n'$. For example, in an agricultural survey, one can select a large sample $s'$ to acquire information about the size of the farm $x$ (auxiliary variable) only, and on the basis of the observed $x$ values, the farms can be classified into small, medium, large, or very large strata.

### 10.3.1 Estimation of Mean and Variance

Let $y_{hi}$ and $x_{hi}$ be the value of the study ($y$) and auxiliary variable ($x$) for the $i$th unit of the $h$th stratum $i = 1,\ldots, N_h$, $h = 1,\ldots, H$. Here the strata sizes $N_h$'s and weights $W_h = N_h/N$'s are not known. Let $\overline{Y}_h = \sum_{i=1}^{N_h} y_{hi}/N_h$ be the population mean for the $h$th stratum and $\overline{Y} = \sum_{h=1}^{H} W_h \overline{Y}_h$ be the mean for the entire population.

Theorem 10.3.1

Let $w_h = n'_h/n'$, then

(i) $E(w_h) = W_h$, (ii) $V(w_h) = g\, W_h(1 - W_h)/n'$, and

(iii) $Cov(w_h, w_{h'}) = -gW_h W_{h'}/n'$ for $h \neq h'$

where $g = (N - n')/(N - 1)$.

Proof

For a fixed $\sum_{k=1}^{H} n'_k = n'$, $\mathbf{n}' = \left(n'_1, \ldots, n'_k, \ldots, n'_H\right)$ follows the generalized hypergeometric distribution

$$Prob\left(n'_1 = t_1, \ldots, n'_k = t_k, \ldots, n'_H = t_H \middle| \sum_{k=1}^{H} n'_k = n'\right)$$

$$= \frac{\binom{N_1}{t_1} \cdots \binom{N_k}{t_k} \cdots \binom{N_H}{t_H}}{\binom{N}{n'}}; t_k = 0, 1, \ldots, \min(n', N_k)$$

For this generalized hypergeometric distribution,

$$E\left(n'_k\right) = n' W_k,\ V\left(n'_k\right) = n' g W_k (1 - W_k) \text{ and } Cov\left(n'_k, n'_{k'}\right) = -n' g W_k W_{k'}$$

Theorem follows from the aforementioned results.

Finally using Rao (1973), we have the following theorem.

Theorem 10.3.2

(i) The estimator $\widehat{\overline{Y}}_{stt} = \sum_{h=1}^{H} w_h \overline{y}(s_h)$ is unbiased $\overline{Y}$

where $\overline{y}(s_h) = \sum_{j \in s_h} y_{hj} / n_h$

(ii) The variance of $\widehat{\overline{Y}}_{stt}$ is

$$V\left(\widehat{\overline{Y}}_{stt}\right) = \frac{1}{n'} \sum_{h=1}^{H} W_h \left\{ \left(\frac{1}{\gamma_h} - 1\right) + g\left(1 - \frac{1}{N_h}\right) \right\} S_{h\gamma}^2$$

$$+ \frac{g}{n'} \sum_{h=1}^{H} W_h \left(\overline{Y}_h - \overline{Y}\right)^2$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) S_\gamma^2 + \sum_{h=1}^{H} \frac{W_h}{n'} \left(\frac{1}{\gamma_h} - 1\right) S_{h\gamma}^2$$

where $(N-1) S_\gamma^2 = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(y_{hi} - \overline{Y}\right)^2$ and $(N_h - 1) S_{h\gamma}^2 =$

$$\sum_{i=1}^{N_h} \left(y_{hi} - \overline{Y}_h\right)^2$$

(iii) A nonnegative unbiased estimator of $V\left(\widehat{\overline{Y}}_{stt}\right)$ is

$$\widehat{V}\left(\widehat{\overline{Y}}_{stt}\right) = \frac{1}{Nn'}\left[\frac{N-1}{n'-1}\sum_{h=1}^{H} n'_h\left(\frac{1}{\gamma_h}-1\right)s^2_{hy} + \frac{N-n'}{n'-1}\left(\sum_{h=1}^{H}\frac{1}{\gamma_h}\sum_{i\in s_h}\gamma^2_{hi} - n'\widehat{\overline{Y}}^2_{stt}\right)\right]$$

$$= \frac{N-1}{N}\sum_{h=1}^{H}\left(\frac{n'_h-1}{n'-1}-\frac{n_h-1}{N-1}\right)\frac{w_h s^2_{hy}}{n_h} + \frac{N-n'}{N(n'-1)}\sum_{h=1}^{H} w_h\left(\overline{y}(s_h)-\widehat{\overline{Y}}_{stt}\right)^2$$

where $g = \dfrac{N-n'}{N-1}$ and $(n_h-1)s^2_{hy} = \sum_{i\in s_h}\left\{\gamma_{hi}-\overline{y}(s_h)\right\}^2$.

Proof

Let $\mathbf{n}' = \left(n'_1, \ldots, n'_h, \ldots, n'_H\right)$. Then,

(i)
$$E\left(\widehat{\overline{Y}}_{stt}\right) = E\left[E\left(\sum_{h=1}^{H} w_h\,\overline{y}(s_h)|\mathbf{n}'\right)\right]$$

$$= E\left[\sum_{h=1}^{H} w_h\,E(\overline{y}(s_h)|\mathbf{n}')\right]$$

$$= E\left[\sum_{h=1}^{H} w_h\ \overline{Y}_h\right]$$

$$= \sum_{h=1}^{H}\{E(w_h)\}\ \overline{Y}_h$$

$$= \sum_{h=1}^{H} W_h\ \overline{Y}_h \text{ (using Theorem 10.3.1)}$$

$$= \overline{Y}$$

(ii)
$$V\left(\widehat{\overline{Y}}_{stt}\right) = E\left[V\left(\sum_{h=1}^{H} w_h\overline{y}(s_h)|\mathbf{n}'\right)\right] + V\left[E\left(\sum_{h=1}^{H} w_h\overline{y}(s_h)|\mathbf{n}'\right)\right] \tag{10.3.1}$$

Now,

$$E\left[V\left(\sum_{h=1}^{H} w_h \overline{y}(s_h)|\mathbf{n}'\right)\right] = E\left[\sum_{h=1}^{H} w_h^2\ V\{\overline{y}(s_h)|\mathbf{n}'\}\right.$$

$$\left. + \sum_{h\neq}^{H}\sum_{h'=1}^{H} w_h w_{h'}\, Cov\{\overline{y}(s_h), \overline{y}(s_{h'})|\mathbf{n}'\}\right]$$

$$= E\left[\sum_{h=1}^{H} w_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{hy}^2\right] \tag{10.3.2}$$

because $V\{\overline{y}(s_h|\mathbf{n}')\} = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{hy}^2$ and $Cov\{\overline{y}(s_h), \overline{y}(s_{h'})|\mathbf{n}'\} = 0$ as the samples $s_h$ and $s_{h'}$ are selected independently for $h \neq h'$.

Writing $n_h = \gamma_h n_h'$ in Eq. (10.3.2) we get

$$E\left[V\left(\sum_{h=1}^{H} w_h\ \overline{y}(s_h)|\mathbf{n}'\right)\right] = \sum_{h=1}^{H}\left(\frac{E(w_h)}{n'\gamma_h} - \frac{E\left(w_h^2\right)}{N_h}\right) S_{hy}^2$$

$$= \sum_{h=1}^{H} W_h \left\{\left(\frac{1}{n'\gamma_h} - \frac{1}{N}\right) - \frac{g}{n'}\left(\frac{1 - W_h}{N_h}\right)\right\} S_{hy}^2 \tag{10.3.3}$$

(using Theorem 10.3.1.)

$$= \frac{1}{n'}\sum_{h=1}^{H} W_h \left\{\left(\frac{1}{\gamma_h} - 1\right) + g\left(1 - \frac{1}{N_h}\right)\right\} S_{hy}^2$$

Furthermore,

$$V\left[E\left(\sum_{h=1}^{H} w_h\ \overline{y}(s_h)|\mathbf{n}'\right)\right] = V\left(\sum_{h=1}^{H} w_h \overline{Y}_h\right)$$

$$= \frac{g}{n'}\left(\sum_{h=1}^{H} W_h(1 - W_h)\overline{Y}_h^2 - \sum_{h\neq}^{H}\sum_{h'}^{H} W_h W_{h'} \overline{Y}_h \overline{Y}_{h'}\right)$$

$$= \frac{g}{n'}\sum_{h=1}^{H} W_h\left(\overline{Y}_h - \overline{Y}\right)^2 \tag{10.3.4}$$

Substituting Eqs. (10.3.3) and (10.3.4) in Eq. (10.3.1) and noting $(N-1)S_y^2 = \sum_{h=1}^{H}(N_h - 1)S_{hy}^2 + \sum_{h=1}^{H} N_h\left(\overline{Y}_h - \overline{Y}\right)^2$ we can verify (ii) of the theorem.

(iii) Now, noting $\frac{1}{N-1}\left[N\sum_{h=1}^{H}\frac{w_h}{n_h}\sum_{i\in s_h}y_{hi}^2 - N\widehat{\overline{Y}}_{stt}^2\left\{-\widehat{V}\left(\widehat{\overline{Y}}_{stt}\right)\right\}\right]$ and $\sum_{h=1}^{H}\frac{w_h}{n'}\left(\frac{1}{\gamma_h}-1\right)s_{hy}^2$ are unbiased estimators of $S_y^2$ and $\sum_{h=1}^{H}\frac{W_h}{n'}\left(\frac{1}{\gamma_h}-1\right)S_h^2$, respectively, we find an unbiased estimator of $V\left(\widehat{\overline{Y}}_{stt}\right)$ from the equation

$$\widehat{V}\left(\widehat{\overline{Y}}_{stt}\right) = \left(\frac{1}{n'}-\frac{1}{N}\right)\frac{1}{N-1}\left[N\sum_{h=1}^{H}\frac{w_h}{n_h}\sum_{i\in s_h}y_{hi}^2 - N\left\{\widehat{\overline{Y}}_{stt}^2 - \widehat{V}\left(\widehat{\overline{Y}}_{stt}\right)\right\}\right] + \sum_{h=1}^{H}\frac{w_h}{n'}\left(\frac{1}{\gamma_h}-1\right)s_{hy}^2 \tag{10.3.5}$$

Eq. (10.3.5) yields

$$\text{i.e.,}\quad \widehat{V}\left(\widehat{\overline{Y}}_{stt}\right) = \frac{1}{Nn'}\left[\frac{N-1}{n'-1}\sum_{h=1}^{H}n'_h\left(\frac{1}{\gamma_h}-1\right)s_{hy}^2 + \frac{N-n'}{n'-1}\left(\sum_{h=1}^{H}\frac{1}{\gamma_h}\sum_{i\in s_h}y_{hi}^2 - n'\widehat{\overline{Y}}_{stt}^2\right)\right] \tag{10.3.6}$$

Furthermore, writing $\sum_{i\in s_h}y_{hi}^2 = (n_h-1)s_{hy}^2 + n_h\{\overline{y}(s_h)\}^2$ in Eq. (10.3.6) we find

$$\widehat{V}\left(\widehat{\overline{Y}}_{stt}\right) = \frac{N-1}{N}\sum_{h=1}^{H}\left(\frac{n'_h-1}{n'-1}-\frac{n_h-1}{N-1}\right)\frac{w_h s_{hy}^2}{n_h} + \frac{N-n'}{N(n'-1)}\sum_{h=1}^{H}w_h\left(\overline{y}(s_h)-\widehat{\overline{Y}}_{stt}\right)^2$$

## 10.3.2 Proportional Allocation

In case the samples are selected from each of the strata by proportional allocation, keeping the total sample size for the second phase fixed as $n$, then $\gamma_h$ becomes equal to $\gamma = n/n'$. Furthermore, if the stratum sizes are large for each of the strata, we derive the following result straight from the Theorem 10.3.2.

Theorem 10.3.3

For proportional allocation with large strata sizes

$$\text{(i)}\ V\left(\overline{Y}_{stt}\right) \cong \sum_{h=1}^{H} W_h \left\{ \frac{(1-f_2)}{n} + \frac{(1-f_1)}{n'} \right\} S_{hy}^2 + \frac{(1-f_1)}{n'} \sum_{h=1}^{H} W_h \left(\overline{Y}_h - \overline{Y}\right)^2$$

and

(ii) An unbiased estimator of $V\left(\widehat{\overline{Y}}_{stt}\right)$ is

$$\widehat{V}\left(\widehat{Y}\right) = \frac{(1-f_1)}{n'} \sum_{h=1}^{H} w_h \left(\overline{y}(s_h) - \widehat{\overline{Y}}_{stt}\right)^2 + \frac{1}{n} \sum_{h=1}^{H} (1-f) w_h s_{hy}^2$$

where $f = n/N$, $f_1 = n'/N$ and $f_2 = n/n'$.

## 10.3.3 Estimation of Proportion

To estimate the proportion of persons who possess some attribute $A$, such as being a smoker or wealthy, we define $y_{hi} = 1$ if the $i$th unit of the $h$th stratum belongs to $A$ and $y_{hi} = 0$ otherwise. Then $\overline{Y}$ equals to $\pi = \sum_{h=1}^{H} W_h \pi_h$, the proportion of persons belonging to the group $A$, and $\pi_h$ becomes the proportion in the $h$th stratum. On substituting $\overline{y}(s_h) = \sum_{j \in s_h} y_{hj} / n_h = \widehat{\pi}_h$ sample proportion for the $h$th stratum, $S_{hy}^2 = \frac{N_h \pi_h (1-\pi_h)}{N_h - 1}$ and $s_{hy}^2 = \frac{n_h \widehat{\pi}_h (1-\widehat{\pi}_h)}{n_h - 1}$ in Theorem 10.3.2, we get the following.

Theorem 10.3.4

(i) An unbiased estimator of $\pi$ is $\widehat{\pi} = \sum_{h=1}^{H} w_h \widehat{\pi}_h$

(ii) The variance of $\widehat{\pi}$ is

$$V(\widehat{\pi}) = \frac{1}{n'} \sum_{h=1}^{H} W_h \left\{ \left( \frac{1}{\gamma_h} - 1 \right) + g \left( 1 - \frac{1}{N_h} \right) \right\} \frac{N_h \pi_h (1-\pi_h)}{N_h - 1} + \frac{g}{n'} \sum_{h=1}^{H} W_h (\pi_h - \pi)^2$$

(iii) An unbiased estimator of $V(\widehat{\pi})$ is

$$\widehat{V}(\widehat{\pi}) = \frac{N-1}{N} \sum_{h=1}^{H} \left( \frac{n_h' - 1}{n' - 1} - \frac{n_h - 1}{N - 1} \right) \frac{w_h \widehat{\pi}_h (1-\widehat{\pi}_h)}{n_h - 1} + \frac{N - n'}{N(n'-1)} \sum_{h=1}^{H} w_h (\widehat{\pi}_h - \widehat{\pi})^2$$

### 10.3.4 Optimum Allocation of Sample Sizes

Following Rao (1973), we consider the simple cost function

$$C = n'c' + \sum_{h=1}^{H} n_h c_h$$

where $c'$ = cost of surveying a unit in the first phase and $c_h$ = cost of surveying a unit of the $h$th stratum in the second phase. Clearly, $c_h$'s are expected to be much larger than $c'$'s. Because $C$ is a random variable, we determine the optimum values of $\gamma_h$'s, which minimize the variance $V\left(\widehat{\overline{Y}}_{stt}\right)$, keeping the expected cost $E(C) = n'c' + n' \sum_{h=1}^{H} \gamma_h W_h c_h$ to a certain level $C^*$.

$$\text{i.e.,} \quad C^* = n'\left(c' + \sum_{h=1}^{H} \gamma_h W_h c_h\right) \tag{10.3.7}$$

Here we perform minimization in two stages. In the first sage, we minimize

$$V\left(\widehat{\overline{Y}}_{stt}\right) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \sum_{h=1}^{H} \frac{W_h}{n'}\left(\frac{1}{\gamma_h} - 1\right) S_{h\gamma}^2 \tag{10.3.8}$$

subject to Eq. (10.3.8) keeping $n'$ as fixed. For minimization consider

$$\phi = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \sum_{h=1}^{H} \frac{W_h}{n'}\left(\frac{1}{\gamma_h} - 1\right) S_{h\gamma}^2 - \lambda\left(C^* - n'c' - n' \sum_{h=1}^{H} \gamma_h W_h c_h\right)$$

where $\lambda$ is an undetermined Lagrange multiplier.

$\frac{\partial \phi}{\partial \gamma_h} = 0$ implies

$$\gamma_h = \frac{1}{n'\sqrt{\lambda}} \frac{S_{h\gamma}}{\sqrt{c_h}} \tag{10.3.9}$$

Eqs. (10.3.7) and (10.3.9) yield

$$\gamma_h = \frac{C^* - n'c'}{n'\left(\sum_{h=1}^{H} W_h S_{h\gamma} \sqrt{c_h}\right)} \frac{S_{h\gamma}}{\sqrt{c_h}} \tag{10.3.10}$$

Finally, putting Eq. (10.3.10) in Eq. (10.3.8) and minimizing it with respect to $n'$, the optimum value of $n'$ comes out as

$$n' = \frac{C^*}{\sqrt{c'}}\left[\sqrt{c'} + \frac{\sum_{h=1}^{H} W_h S_{h\gamma}\sqrt{c_h}}{\sqrt{S_\gamma^2 - \sum_{h=1}^{H} W_h S_{h\gamma}^2}}\right]^{-1} \tag{10.3.11}$$

Substituting Eq. (10.3.11) in Eq. (10.3.10) the optimum value of $\gamma_h$ is obtained as

$$\gamma_h = \gamma_{h0} = S_{h\gamma}\sqrt{\frac{c'}{c_h}}\Bigg/\sqrt{S_\gamma^2 - \sum_{h=1}^{H} W_h S_{h\gamma}^2} \tag{10.3.12}$$

provided $\gamma_h \leq 1$ for all $h = 1,\ldots, H$. It should be noted that $\gamma_h > 0$ because $\left(S_\gamma^2 - \sum_{h=1}^{H} W_h S_{h\gamma}^2\right) > 0$ except in pathological situations. If some $\gamma_h$ exceed 1, we should set them the value at 1 and determine the rest $\gamma_h$'s so that all of them are less than or equal to 1.

Now, substituting the optimum values of $n'$ and $\gamma_h$ from Eqs. (10.3.11) and (10.3.12) in the expression of $V\left(\widehat{\overline{Y}}_{stt}\right)$, we get the minimum value of $V\left(\widehat{\overline{Y}}_{stt}\right)$ as

$$V_{\min}\left(\widehat{\overline{Y}}_{stt}\right) = \frac{1}{C^*}\left[\sum_{h=1}^{H} W_h S_{h\gamma}\sqrt{c_h} + \sqrt{\left(S_\gamma^2 - \sum_{h=1}^{H} W_h S_{h\gamma}^2\right)c'}\right]^2 - \frac{S_\gamma^2}{N} \tag{10.3.13}$$

Example 10.3.1

A pilot sample of 1000 households was selected from 10,000 households in a certain region by SRSWOR method. On the basis of the pilot survey report, the selected households were classified into three categories: The high, middle, and low income groups. From each of the income groups subsamples were selected by SRSWOR method. The sample mean, sample

variance of the monthly income, and sample proportion of unemployed persons for the final survey are presented in the following table.

| Income group | Pilot survey | Final survey | | | |
|---|---|---|---|---|---|
| | No. of households | No. of households | Sample mean (monthly income in $) | Sample variance (monthly income) | Proportion of unemployed persons |
| $h$ | $n'_h$ | $n_h$ | $\overline{y}(s_h)$ | $s^2_{hy}$ | $\widehat{\pi}_h$ |
| Low | 325 | 100 | 1575 | 4000 | 0.75 |
| Middle | 475 | 75 | 5750 | 3500 | 0.40 |
| High | 200 | 75 | 8000 | 2500 | 0.20 |

Estimate the average income and proportion of unemployed persons in the region along with their standard error.

**Estimation of average income and its standard error**

| $h$ | $w_h = n'_h/n'$ | $w_h\overline{y}(s_h)$ | $w_h\left(\overline{y}(s_h) - \widehat{\overline{Y}}_{stt}\right)^2$ | $\left(\frac{n'_h-1}{n'-1} - \frac{n_h-1}{N-1}\right)\frac{w_h s^2_{hy}}{n_h}$ |
|---|---|---|---|---|
| 1 | 0.325 | 511.875 | 3,471,208.330 | 4.087 |
| 2 | 0.475 | 2731.250 | 390,650.576 | 10.353 |
| 3 | 0.200 | 1600.000 | 1,993,171.953 | 1.278 |
| Total | 1.000 | 4843.125 | 5,855,030.859 | 15.719 |

$$\text{Estimated average monthly income} = \widehat{\overline{Y}}_{stt} = \sum_{h=1}^{H} w_h\,\overline{y}(s_h) = \$4843.125.$$

$$\text{Estimated variance of } \widehat{\overline{Y}}_{stt} = \widehat{V}\left(\widehat{\overline{Y}}_{stt}\right) \cong \sum_{h=1}^{H}\left(\frac{n'_h-1}{n'-1} - \frac{n_h-1}{N-1}\right)\frac{w_h s^2_{hy}}{n_h} + \frac{N-n'}{N(n'-1)}\sum_{h=1}^{H} w_h\left(\overline{y}(s_h) - \widehat{\overline{Y}}_{stt}\right)^2$$

$$= 15.719 + \{(10,000-1000)/(10,000\times 999)\} \times 5855030.859$$

$$= \$5290.521$$

$$\text{Estimated standard error of } \widehat{\overline{Y}}_{stt} = \sqrt{\widehat{V}\left(\widehat{\overline{Y}}_{stt}\right)} = \$72.74$$

**Estimation of proportion of unemployed persons and its standard error**

| $h$ | $w_h = n_h'/n'$ | $w_h\widehat{\pi}_h$ | $w_h(\widehat{\pi}_h - \widehat{\pi})^2$ | $\left(\frac{n_h'-1}{n'-1} - \frac{n_h-1}{N-1}\right)\frac{w_h\widehat{\pi}_h(1-\widehat{\pi}_h)}{n_h-1}$ |
|---|---|---|---|---|
| 1 | 0.325 | 0.244 | 0.0248 | 0.00019 |
| 2 | 0.475 | 0.190 | 0.0025 | 0.00072 |
| 3 | 0.200 | 0.040 | 0.0149 | 0.00008 |
| Total | 1.000 | 0.474 | 0.0423 | 0.00099 |

$$\text{Estimated proportion of unemployed persons} = \widehat{\pi} = \sum_{h=1}^{H} w_h\widehat{\pi}_h = 0.4737$$

$$\text{Estimated variance of } \widehat{\pi} = \widehat{V}(\widehat{\pi}) \cong \sum_{h=1}^{H}\left(\frac{n_h'-1}{n'-1} - \frac{n_h-1}{N-1}\right)\frac{w_h\widehat{\pi}_h(1-\widehat{\pi}_h)}{n_h-1} + \frac{N-n'}{N(n'-1)}\sum_{h=1}^{H} w_h(\widehat{\pi}_h - \widehat{\pi})^2$$

$$= 0.00099 + \{(10000-1000)/(10000\times 999)\}\times 0.0423 = 0.0010$$

$$\text{Estimated standard error of } \widehat{\pi} = \sqrt{\widehat{V}(\widehat{\pi})} = 0.032$$

## 10.4 TWO-PHASE SAMPLING FOR SELECTION OF SAMPLE

In case the auxiliary information ($x$) selected in the first phase is positive, it can be used as a measure of size for the selection of sample in the second phase. In this section we describe the use of PPSWR and Rao–Hartley–Cochran (1962, RHC) methods of sampling for the selection of samples in the second phase when the first-phase sample is selected by SRSWOR method.

### 10.4.1 Probability Proportional to Size With Replacement Sampling

Let us suppose that in the first phase, a sample $s'$ of size $n'$ be selected by SRSWOR method and the values of the auxiliary variable $x$ are obtained. In the second phase, a subsample $s$ of size $n$ is selected from $s'$ by using PPSWR method using $x$'s as measure of size, i.e., the normed size measure for the $i$th unit is $p_i' = x_i/X'$ for $i \in s'$, where $X' = \sum_{i\in s'} x_i$. The estimator for the population total $Y$ is given by

$$\widehat{Y}_{ppst} = \frac{N}{n'}\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}\right) \tag{10.4.1}$$

Theorem 10.4.1

(i) $\widehat{Y}_{ppst}$ is an unbiased estimator for $Y$.

(ii) Variance of $\widehat{Y}_{ppst}$ is

$$V\left(\widehat{Y}_{ppst}\right) = N^2\left[\frac{(n'-1)}{N(N-1)n'n}\sum_{i\in U} p_i\left(\frac{y_i}{p_i} - Y\right)^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2\right]$$

where $p_i = x_i/X$ and $X = \sum_{i\in U} x_i$.

(iii) An unbiased estimator of $V\left(\widehat{Y}_{ppst}\right)$ is

$$\widehat{V}\left(\widehat{Y}_{ppst}\right) = N^2\left[\frac{N-1}{N(n'-1)n'}\,\widehat{V}\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right) + \frac{N-n'}{N(n'-1)}\left(\frac{1}{n'n}\sum_{i\in s}\frac{y_i^2}{p_i'} - \frac{\widehat{Y}_{ppst}^2}{N^2}\right)\right]$$

where $\widehat{V}\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right) = \frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{y_i}{p_i'} - \frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}\right)^2$.

Proof

(i)
$$\begin{aligned} E\left(\widehat{Y}_{ppst}\right) &= E\left[\frac{N}{n'}E\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right)\right] \\ &= NE\left(\frac{1}{n'}\sum_{i\in s'} y_i\right) \\ &= NE(\overline{y}(s')) = Y \end{aligned}$$

(ii)
$$\begin{aligned} V\left(\widehat{Y}_{ppst}\right) &= E\left[\frac{N^2}{n'^2}V\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right)\right] + V\left[\frac{N}{n'}E\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right)\right] \\ &= E\left[\frac{N^2}{n'^2}\left\{\frac{1}{2n}\sum_{i\neq}\sum_{j\in s'} p_i'p_j'\left(\frac{y_i}{p_i'} - \frac{y_i}{p_i'}\right)^2\right\}\right] + V(N\,\overline{y}(s')) \\ &= \frac{N^2}{n'^2}\frac{1}{2n}\sum_{i\neq}\sum_{j\in U} p_ip_j\left(\frac{y_i}{p_i} - \frac{y_i}{p_i}\right)^2\pi_{ij}^0 + N^2\left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 \end{aligned}$$

[where $\pi_{ij}^0$ = inclusion probability of $i$th and $j$th unit $(i \neq j)$ in $s' = n'(n'-1)/\{N(N-1)\}$]

$$= N^2\left[\frac{(n'-1)}{n'nN(N-1)}\sum_{i\in U} p_j\left(\frac{y_i}{p_i} - Y\right)^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2\right]$$

(iii) Unbiased estimators of $V\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right)$ and $S_y^2$ are given, respectively, by

$$\widehat{V}\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right)=\frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{y_i}{p_i'}-\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}\right)^2$$

and

$$\widehat{S}_y^2=\frac{1}{(N-1)}\left\{\frac{N}{n'n}\sum_{i\in s}\frac{y_i^2}{p_i'}-\frac{\widehat{Y}_{ppst}^2-\widehat{V}\left(\widehat{Y}_{ppst}\right)}{N}\right\}$$

Then an unbiased estimator of $V\left(\widehat{Y}_{ppst}\right)$ is obtained from the following equation

$$\begin{aligned}\widehat{V}\left(\widehat{Y}_{ppst}\right)=&N^2\left[\frac{1}{n'^2}\widehat{V}\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right)+\left(\frac{1}{n'}-\frac{1}{N}\right)\widehat{S}_y^2\right]\\=&N^2\left[\frac{1}{n'^2}\widehat{V}\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right)\right.\\&\left.+\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{1}{(N-1)}\left\{\frac{N}{n'}\frac{1}{n}\sum_{i\in s}\frac{y_i^2}{p_i'}-\frac{\widehat{Y}_{ppst}^2-\widehat{V}\left(\widehat{Y}_{ppst}\right)}{N}\right\}\right]\end{aligned}\tag{10.4.2}$$

Finally, Eq. (10.4.2) yields

$$\begin{aligned}\widehat{V}\left(\widehat{Y}_{ppst}\right)=N^2\left[\frac{N-1}{N(n'-1)n'}\widehat{V}\left(\frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i'}|s'\right)\right.\\\left.+\frac{N-n'}{N(n'-1)}\left(\frac{1}{n'n}\sum_{i\in s}\frac{y_i^2}{p_i'}-\frac{\widehat{Y}_{ppst}^2}{N^2}\right)\right]\end{aligned}$$

### 10.4.2 Rao–Hartley–Cochran Sampling

Avadhani and Sukhatme (1970) considered two-phase sampling, where in the first phase, a sample $s'$ of size $n'$ is selected by SRSWOR method and in the second phase, a subsample $s$ of size $n$ is selected from $s'$ by using the RHC method (see Section 5.6) using normed size measure $p_i'=x_i/X'$ for the $i$th unit, $i\in s'$. Let $P_i'$ be the sum of the $p'$ values of the group

containing the $i$th unit in selecting the subsample $s$ by RHC scheme from $s'$. The proposed estimator for $Y$ is

$$\widehat{Y}_{rhct} = \frac{N}{n'} \sum_{j \in s} \frac{y_j}{p'_j} P'_i \tag{10.4.3}$$

Theorem 10.4.2

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

(ii) $V\left(\widehat{Y}_{rhct}\right) = N^2 \left[ \frac{(n'-n)}{n'nN(N-1)} \sum_{i \in U} p_i \left( \frac{y_i}{p_i} - Y \right)^2 + \left( \frac{1}{n'} - \frac{1}{N} \right) S_y^2 \right]$

(iii) $\widehat{V}\left(\widehat{Y}_{rhct}\right) = N^2 \left[ \frac{N-1}{n'N(n'-1)} \widehat{V} \left( \sum_{j \in s} \frac{y_j}{p'_j} P'_j | s' \right) + \frac{N-n'}{N(n'-1)} \left( \frac{1}{n'} \sum_{i \in s} \frac{y_i^2}{p'_i} P'_i - \frac{\widehat{Y}_{rhct}^2}{N^2} \right) \right]$

where $\widehat{V} \left( \sum_{j \in s} \frac{y_j}{p'_j} P'_j | s' \right) = \frac{n'-n}{n'(n-1)} \sum_{j \in s} P'_j \left( \frac{y_j}{p'_j} - \sum_{j \in s} \frac{y_j}{p'_j} P'_j \right)^2$.

Proof

(i) $E\left(\widehat{Y}_{rhct}\right) = E \left[ \frac{N}{n'} E \left( \sum_{j \in s} \frac{y_j}{p'_j} P'_i | s' \right) \right]$

$= E \left( \frac{N}{n'} \sum_{j \in s'} y_j \right)$

$= Y$

(ii) $V\left(\widehat{Y}_{rhct}\right) = E \left[ \frac{N^2}{n'^2} V \left( \sum_{j \in s} \frac{y_j}{p'_j} P'_i | s' \right) \right] + V \left[ \frac{N}{n'} E \left( \sum_{j \in s} \frac{y_j}{p'_j} P'_i | s' \right) \right]$

$= \frac{N^2}{n'^2} \frac{n'-n}{n(n'-1)} \frac{1}{2} E \sum_{i \neq} \sum_{j \in s'} p_i p_j \left( \frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 + N^2 V \left( \frac{1}{n'} \sum_{j \in s'} y_j \right)$

$= N^2 \left[ \frac{(n'-n)}{n'nN(N-1)} \sum_{i \in U} p_i \left( \frac{y_i}{p_i} - Y \right)^2 + \left( \frac{1}{n'} - \frac{1}{N} \right) S_y^2 \right]$

(iii) Now, noting $\widehat{V}\left(\sum_{j\in s}\frac{y_j}{p'_j}P'_j|s'\right) = \frac{n'-n}{n'(n-1)}\sum_{j\in s}P'_j\left(\frac{y_j}{p'_j}-\sum_{j\in s}\frac{y_j}{p'_j}P'_j\right)^2$

and

$\widehat{S}^2_{rhct} = \frac{1}{N-1}\left\{\frac{N}{n'}\sum_{i\in s}\frac{y_i^2}{p'_i}P'_i - \frac{\widehat{Y}^2_{rhct}-\widehat{V}\left(\widehat{Y}_{rhct}\right)}{N}\right\}$ are unbiased estimators of $V\left(\sum_{j\in s}\frac{y_j}{p'_j}P'_j|s'\right)$ and $S_y^2$, respectively, we find $\widehat{V}\left(\widehat{Y}_{rhct}\right)$ as an unbiased estimator of $V\left(\widehat{Y}_{rhct}\right)$ from the following equation:

$$\widehat{V}\left(\widehat{Y}_{rhct}\right) = \frac{N^2}{n'^2}\widehat{V}\left(\sum_{j\in s}\frac{y_j}{p'_j}P'_j|s'\right) + N^2\left(\frac{1}{n'}-\frac{1}{N}\right)\widehat{S}^2_{rhct}$$

$$\text{i.e.,}\quad \widehat{V}\left(\widehat{Y}_{rhct}\right) = N^2\left[\frac{1}{n'^2}\widehat{V}\left(\sum_{j\in s}\frac{y_j}{p'_j}P'_j|s'\right) + \left(\frac{1}{n'}-\frac{1}{N}\right)\frac{1}{N-1}\left\{\frac{N}{n'}\sum_{i\in s}\frac{y_i^2}{p'_i}P'_j - \frac{\widehat{Y}^2_{rhct}-\widehat{V}\left(\widehat{Y}_{rhct}\right)}{N}\right\}\right]$$

$$\text{i.e.,}\quad \widehat{V}\left(\widehat{Y}_{rhct}\right) = N^2\left[\frac{N-1}{n'N(n'-1)}\widehat{V}\left(\sum_{j\in s}\frac{y_j}{p'_j}P'_j|s'\right) + \frac{N-n'}{N(n'-1)}\left(\frac{1}{n'}\sum_{i\in s}\frac{y_i^2}{p'_i}P'_i - \frac{\widehat{Y}^2_{rhct}}{N^2}\right)\right]$$

## 10.5 TWO-PHASE SAMPLING FOR STRATIFICATION AND SELECTION OF SAMPLE

Arnab (1991) considered a two-phase sampling procedure, where in the first phase, a PPSWR sample $s'$ of size $n'$ is selected from the entire population $U$ with known size measures $z_i$'s ($>0$ for $i = 1,\ldots, N$), and information of the auxiliary variable $x_i$'s ($i \in s'$) is obtained. By using the information of the $x_i$'s gathered in the initial sample $s'$, the $n'$ sampled units are assigned to a predetermined number of $L$ strata. Let $\widetilde{s}_h$ be the sample of size $n'_h$ corresponding to the $h$th stratum $U_h$ ($h = 1,\ldots, L$). In the second phase, subsamples $s_h$'s of sizes $n_h = \gamma_h n'_h$ (assuming integer and $0 \le \gamma_h \le 1$) are selected independently from each of the $h$th stratum by PPSWR method with normed size measure $q_{hi} = (x_{hi}/z_{hi})/\sum_{i\in s'_h} x_{hi}/z_{hi}$ for the $i$th

unit $(i \in \widetilde{s}_h)$, where $z_{hi}$ and $x_{hi}$ are the values of the variables $z$ and $x$ for the $i$th unit of the $h$th stratum, respectively. Here it is assumed that $n'$ is so large that $Prob(n'_h \geq 1) = 1$. The proposed estimator for the population total $Y$ is

$$\widehat{Y}_A = \sum_h w_h \widehat{Y}_{Ah} \tag{10.5.1}$$

where $\widehat{Y}_{Ah} = \frac{1}{n_h} \sum_{j \in s_h} \frac{y_{hj} - \lambda_h x_{hj}}{\left(n'_h p_{hj}\right) q_{hj}} + \lambda_h \frac{1}{n'_h} \sum_{j \in \widetilde{s}_h} \frac{x_{hj}}{p_{hj}}, \qquad p_{hj} = z_{hj}/Z,$

$Z = \sum_{i \in U} z_i$, $\lambda_h$ is a known constant, and $\sum_{j \in s}$ denotes sum over the units in $s$ with repetition.

Theorem 10.5.1

(i) $E(\widehat{Y}_A) = Y$

(ii) $V(\widehat{Y}_A) = \frac{1}{n'} \left[ \sum_h \left(1 - \frac{1}{n' P_h}\right) \frac{\sigma_h^2(d_{\lambda h}|x)}{P_h \gamma_h} + \sigma^2(y|z) \right]$

where $w_h = n'_h/n'$, $P_h = Z_h/Z$, $Z_h = \sum_{i \in U_h} z_{hi}$, $\sigma_h^2(d_{\lambda h}|x) = \sum_{j \in U_h} \frac{d_{\lambda hj}^2}{x_{hi}} X_h - D_{\lambda h}^2$,

$X_h = \sum_{i \in U_h} x_{hi}, \qquad d_{\lambda hj} = y_{hj} - \lambda_h x_{hj}, \qquad D_{\lambda h} = \sum_{j \in U_h} d_{\lambda hj} = Y_h - \lambda_h X_h, \qquad$ and

$\sigma^2(y|z) = \sum_{h=1}^{H} \sum_{j \in U_h} \frac{y_{hj}^2}{p_{hj}} - Y^2.$

Proof

(i) $E(\widehat{Y}_A) = E \sum_h w_h E(\widehat{Y}_{Ah} | s'_h)$

$= E \sum_h w_h \sum_{j \in \widetilde{s}_h} \frac{y_{hj}}{\left(n'_h p_{hj}\right)}$

$= E \sum_h \frac{w_h}{P_h} E \left( \sum_{j \in \widetilde{s}_h} \frac{y_{hj} P_h}{\left(n'_h p_{hj}\right)} | \mathbf{w_h} \right)$

(where $\mathbf{w_h} = (w_1, \ldots, w_L)$)

$= E \sum_h \frac{w_h}{P_h} Y_h$

$= Y$

(noting $E(w_h) = P_h$)

(ii) $V\left(\widehat{Y}_A\right) = E\left[\sum_h w_h^2 V\left(\widehat{Y}_{Ah}|\widetilde{s}_h\right)\right] + V\left[\sum_h w_h E\left(\widehat{Y}_{Ah}|\widetilde{s}_h\right)\right]$

Now,

$$V\left\{\sum_h w_h E\left(\widehat{Y}_{Ah}|\widetilde{s}_h\right)\right\} = V\left\{\sum_h w_h \sum_{j\in\widetilde{s}_h} \frac{\gamma_{hj}}{\left(n'_h p_{hj}\right)}\right\}$$

$$= E\left\{\sum_h \frac{w_h^2}{P_h^2} V\left(\frac{1}{n'_h}\sum_{j\in\widetilde{s}_h}\frac{\gamma_{hj}}{p'_{hj}}|\mathbf{w_h}\right)\right\}$$

$$+ V\left\{\sum_h \frac{w_h}{P_h} E\left(\frac{1}{n'_h}\sum_{j\in\widetilde{s}_h}\frac{\gamma_{hj}}{p'_{hj}}|\mathbf{w_h}\right)\right\}$$

$$\left(\text{where } p'_{hi} = p_{hi}/P_h\right)$$

$$= E\left[\sum_h \frac{w_h^2}{P_h^2}\frac{1}{n'_h}\left(\sum_{j\in U_h}\frac{\gamma_{hj}^2}{p'_{hj}} - Y_h^2\right) + V\left(\sum_h \frac{w_h}{P_h}Y_h\right)\right]$$

$$= \frac{1}{n'}\left[\sum_h \frac{1}{P_h}\left(\sum_{j\in U_h}\frac{\gamma_{hj}^2}{p'_{hj}} - Y_h^2\right) + \left(\sum_h \frac{Y_h^2}{P_h} - Y^2\right)\right]$$

$$= \frac{1}{n'}\left(\sum_{h=1}^{H}\sum_{j\in U_h}\frac{\gamma_{hj}^2}{p_{hj}} - Y^2\right)$$

$$= \frac{\sigma^2(\gamma|z)}{n'} \qquad (10.5.2)$$

and

$$E\left\{\sum_h w_h^2 V\left(\widehat{Y}_{Ah}|\widetilde{s}_h\right)\right\} = E\left\{\sum_h w_h^2 V\left(\sum_{j\in s_h}\frac{\gamma_{hj}-\lambda_h x_{hj}}{\left(n_h' p_{hj}\right)n_h q_{hj}}|\widetilde{s}_h\right)\right\}$$

$$= E\sum_h \frac{w_h^2}{n_h}\left\{\sum_{j\in\widetilde{s}_h}\frac{d_{\lambda hj}^2}{\left(n_h' p_{hj}\right)^2 q_{hj}} - \left(\sum_{j\in\widetilde{s}_h}\frac{d_{\lambda hj}}{n_h' p_{hj}}\right)^2\right\}$$

$$= E\sum_h \frac{w_h^2}{n_h} E\left\{\sum_{j\in\widetilde{s}_h}\frac{d_{\lambda hj}^2}{\left(n_h' p_{hj}\right)^2}\frac{p_{hj}}{q_{hj}'}\sum_{j\in\widetilde{s}_h}\frac{q_{hj}'}{p_{hj}} - \left(\sum_{j\in\widetilde{s}_h}\frac{d_{\lambda hj}}{n_h' p_{hj}}\right)^2 |\mathbf{w_h}\right\}$$

$$\left(\text{where } q_{hj}' = x_{hj}/X_h\right)$$

$$= E\sum_h \frac{w_h^2}{n_h} E\left\{\sum_{j\in\widetilde{s}_h}\frac{d_{\lambda hj}^2}{\left(n_h' p_{hj}\right)^2} + \sum_{j\in\widetilde{s}_h}\sum_{k(\neq j)\in\widetilde{s}_h}\frac{d_{\lambda hj}^2}{\left(n_h' p_{hj}\right)^2}\frac{p_{hj}}{q_{hj}'}\frac{q_{hk}'}{p_{hk}}\right.$$

$$\left. - \left(\sum_{j\in\widetilde{s}_h}\frac{d_{\lambda hj}}{n_h' p_{hj}}\right)^2 |\mathbf{w_h}\right\}$$

$$= E\sum_h \frac{w_h^2}{n_h}\left[\frac{1}{\left(n_h'\right)^2}\left\{\frac{n_h'}{P_h}\sum_{j\in U_h}\frac{d_{\lambda hj}^2}{p_{hj}} + \frac{n_h'\left(n_h'-1\right)}{P_h^2}\left(\sum_{j\in U_h}\frac{d_{\lambda hj}^2}{q_{hj}'}\right)\right\}\right.$$

$$\left. - \frac{1}{P_h^2}\left\{\frac{1}{n_h'}\left(\sum_{j\in U_h}\frac{d_{\lambda hj}^2}{p_{hj}'} - D_{\lambda h}^2\right) + D_{\lambda h}^2\right\}\right]$$

$$\left(\text{noting } \sum_{j\in U_h} q_{hj}' = 1\right)$$

$$= E\sum_h \frac{w_h^2}{n_h}\frac{n_h'-1}{n_h' P_h^2}\left(\sum_{j\in U_h}\frac{d_{\lambda hj}^2}{x_{hj}}X_h - D_{\lambda h}^2\right)$$

$$\left(\text{noting } p_{hj}' = \frac{z_{hj}}{Z_h} = \frac{p_{hj}}{P_h} \text{ and } q_{hj}' = \frac{x_{hj}}{X_h}\right)$$

Finally, writing $n_h = \gamma_h n'_h$, we get

$$E\sum_h w_h^2 V\left(\widehat{Y}_{Ah}|\widetilde{s}_h\right) = \sum_h \left(1 - \frac{1}{n'P_h}\right)\frac{\sigma_h^2(d_{\lambda h}|x)}{n'\gamma_h P_h} \tag{10.5.3}$$

The proof (ii) follows from Eqs. (10.5.2) and (10.5.3).

## 10.6 EXERCISES

**10.6.1** Let an initial sample $s'$ of size $n'$ be selected from a finite population of size $N$ by SRSWOR method and information on the auxiliary variable $x$ is collected. Then an independent sample $s$ of size $n$ is selected by SRSWOR method from the entire population, and information on the auxiliary ($x$) and study variable ($y$) is obtained. Let $t = \overline{y}(s) - k\{\overline{x}(s) - \overline{x}(s')\}$.

Show that

(i) $E(t) = \overline{Y}$,

(ii) $V(t) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_y^2 + k^2 S_x^2 - 2kS_{xy}\right) + k^2\left(\frac{1}{n'} - \frac{1}{N}\right)S_x^2$

(iii) $\widehat{V}(t) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(s_y^2 + k^2 s_x^2 - 2ks_{xy}\right) + k^2\left(\frac{1}{n'} - \frac{1}{N}\right)s_x^2$ is an unbiased estimator for $V(t)$

(iv) Find the value of k for which $V(t)$ attains a minimum (Raj, 1968).

**10.6.2** Initially, a sample $s'$ of size $n'$ is selected from a finite population of size $N$ by PPSWR method using $p_i$ as a normed size measure for the $i$th unit, and information on the auxiliary variable $x$ is collected. Then an independent sample $s$ of size $n$ is selected by PPSWR method using the same normed size measure for the $i$th unit, and information on the auxiliary ($x$) and study variable ($y$) is obtained. Let $t = \frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i} - k\left(\frac{1}{n}\sum_{i\in s}\frac{x_i}{p_i} - \frac{1}{n'}\sum_{i\in s'}\frac{x_i}{p_i}\right)$ and $\sum_{i\in s}$ is the sum over units in $s$ with repetition. Show that

(i) $E(t) = \overline{Y}$

(ii) $V(t) = \frac{\sigma^2(d_k|p)}{n} + k^2\frac{\sigma^2(x|p)}{n'}$, where $\sigma^2(d_k|p) = \sum_{i\in U}\frac{d_{ki}^2}{p_i} - D_k^2$,

$\sigma^2(x|p) = \sum_{i\in U}\frac{x_i^2}{p_i} - X^2$, $d_{ki} = y_i - kx_i$, and $D_k = \sum_{i\in U} d_{ki}$

(iii) $\widehat{V}(t) = \frac{\widehat{\sigma}^2(d_k|p)}{n} + k^2\frac{\widehat{\sigma}^2(x|p)}{n'}$ is an unbiased estimator for $V(t)$, where

$$\widehat{\sigma}^2(d_k|p) = \frac{\sum_{i\in s}\left(\frac{d_{ki}}{p_i} - \frac{1}{n}\sum_{i\in s}\frac{d_{ki}}{p_i}\right)^2}{(n-1)} \text{ and } \widehat{\sigma}^2(x|p) = \frac{\sum_{i\in s'}\left(\frac{x_i}{p_i} - \frac{1}{n'}\sum_{i\in s}\frac{x_i}{p_i}\right)^2}{(n'-1)}$$

(iv) Find the value of k for which $V(t)$ attains the minimum (Raj, 1968).

**10.6.3** In the first phase an initial sample $s'$ of size $n'$ is selected by SRSWOR method, and information on the auxiliary variable $x$ is collected. At the second phase, a subsample $s$ of size $n$ is selected from $s'$ by Lahiri—Midzuno—Sen (1951, 1952, 1953) sampling procedure using $x$ as a measure of size.

Show that $T = N\frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i}\left(\sum_{i\in s'} x_i/n'\right)$ is an unbiased estimator for the population total $Y$. Derive the variance of $T$ and find an unbiased estimator of $V(T)$ (Raj, 1968).

**10.6.4** In the first phase a sample $s'$ of size $n'$ is selected by RHC sampling scheme using a normed size measure $p_i$ attached to the $i$th unit. At the second phase a subsample $s$ of size $n$ is selected from $s'$ by SRSWOR method. Let $T = \frac{n'}{n}\sum_{i\in s}\frac{y_i}{p_i}P_i - k\left(\frac{n'}{n}\sum_{i\in s}\frac{x_i}{p_i}P_i - \sum_{i\in s'}\frac{x_i}{p_i}P_i\right)$, where $P_i$ is the sum of $p_j$ values for the group (which was formed for the selection of the sample) containing the $i$th unit.

**(a)** Verify the following results:

(i) $E(T) = Y$

(ii) $V(T) = \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{1}{(N-1)}\left[n'\overline{V}(d_k) + (N-n')\sigma^2(d_k|p)\right] + \frac{N-n'}{n'(N-1)}\sigma^2(y|p)$

where $d_{ki} = y_i - kx_i$, $\sigma^2(y|p) = \sum_{i\in U}\frac{y_i^2}{p_i} - Y^2$, $\sigma^2(d_k|p) = \sum_{i\in U}\frac{d_{ki}^2}{p_i} - D_k^2$, $D_k = \sum_{i\in U} d_{ki}$, and $\overline{V}(d_k) = N\sum_{i\in U} d_{ki}^2 - D_k^2$.

**(b)** Find an unbiased estimator of $\widehat{V}(T)$ and the optimum value of value of $k$ that minimizes $V(T)$ (Arnab, 1979).

**10.6.5** Let $\pi_{1i}$ and $\pi_{1ij}$ be the first-order and second-order inclusion probabilities for the first-phase sample $s_1$, and $\pi_{2i|s_1}$ and $\pi_{2ij|s_1}$ be the

conditional first-order and second-order inclusion probabilities for the second-phase sample $s_2$ given $s_1$. Show that

(i) $\widehat{Y}_{DE} = \sum_{i \in s_2} \frac{y_i}{\pi_i^*}$ with $\pi_i^* = \pi_{1i}\pi_{2i|s_1}$ is an unbiased estimator of the total $Y$.

(ii) The variance of $\widehat{Y}_{DE}$ can be estimated unbiasedly by using any of the following estimators:

**(a)** $\widehat{V}_1 = \sum_{i \in s_2} \frac{\Delta_{1ij}}{\pi_{ij}^*} \frac{y_i}{\pi_{1i}} \frac{y_j}{\pi_{1j}} + \sum_{i \in s_2} \sum_{j \in s_2} \frac{\Delta_{2ij|s_1}}{\pi_{ij|s_1}} \frac{y_i}{\pi_i^*} \frac{y_j}{\pi_j^*}$

and

**(b)** $\widehat{V}_2 = \sum_{i \in s_2} \sum_{\substack{j \in s_2 \\ i<j}} \frac{\Delta_{1ij}}{\pi_{ij}^*} \left( \frac{y_i}{\pi_{1i}} - \frac{y_j}{\pi_{1j}} \right)^2 + \sum_{i \in s_2} \sum_{\substack{j \in s_2 \\ i<j}} \frac{\Delta_{2ij|s_1}}{\pi_{ij|s_1}} \left( \frac{y_i}{\pi_i^*} - \frac{y_j}{\pi_j^*} \right)^2$

where $\pi_{ij}^* = \pi_{1ij}\pi_{2ij|s_1}$, $\Delta_{1ij} = \pi_{1i}\pi_{ij} - \pi_{1ij}$, and $\Delta_{2ij|s_1} = \pi_{2i|s_1}\pi_{2j|s_1} - \pi_{2ij|s_1}$ (Haziza et al., 2011).

**10.6.6** To estimate the average price of a two-bedroom flat in a locality, a sample of 75 two-bedroom flats was selected by SRSWOR from the locality of 790 flats and eye estimates of the price of the selected flats were obtained. In phase 2, a random sample of 15 flats was selected from the selected 75 flats by SRSWOR method and the prices of the flats were determined by an appropriate evaluation agency. The findings are given in the following table:

| Serial number of flats | Price of the flat by eye estimation (in 000 $) (x) | Price of the flat by valuation agency (in 000 $) (y) |
|---|---|---|
| 1 | 450 | 550 |
| 2 | 565 | 600 |
| 3 | 700 | 650 |
| 4 | 850 | 900 |
| 5 | 650 | 750 |
| 6 | 700 | 820 |
| 7 | 850 | 970 |
| 8 | 550 | 700 |
| 9 | 600 | 650 |
| 10 | 750 | 700 |
| 11 | 900 | 850 |
| 12 | 850 | 950 |
| 13 | 840 | 900 |
| 14 | 700 | 850 |
| 15 | 760 | 800 |

Estimate the average price of the flat along with its standard error. If the cost evaluation of a flat by an agency is five times that of the eye estimation, determine the proportion of the second-phase sample that should be selected from the first-phase sample.

**10.6.7** To estimate the average daily wages of a factory of 800 workers, an initial sample of 125 workers was selected at random by SRSWOR method. The workers were then stratified according to their skill. Finally, samples of workers were selected from each of the strata by SRSWOR method, and information regarding the daily wages was obtained. The data are given below.

(i) Give an estimate of the average daily wages of the factory workers.

(ii) Estimate the standard error of your estimator.

| Workers type | No. of workers in the initial sample | No. of workers in the final sample | Daily wages (in $) |
|---|---|---|---|
| Highly skilled | 10 | 5 | 625, 750, 1025, 1500, 1000 |
| Skilled | 60 | 10 | 320, 350, 375, 200, 300, 325, 325, 475, 350, 300 |
| Semiskilled | 40 | 8 | 200, 175, 150, 240, 250, 180, 120, 170 |
| Unskilled | 20 | 8 | 150, 175, 100, 140, 200, 160, 100, 100 |