CHAPTER 11

Repetitive Sampling

11.1 INTRODUCTION

In a repetitive or successive sampling, we survey the same population on different occasions for the purpose of estimating the population parameters for the most recent occasion, for changes over previous occasions, estimating parametric functions over last few occasions, among others. For example, several government agencies collect unemployment data on a monthly basis to estimate the rate of unemployment for the most recent month, change over previous month, average of unemployment rate for the last one year, etc. In repetitive sampling, one has to effectively utilize the accumulated data procured in the course of the survey along with other auxiliary information that may also incidentally be available. The objectives of repetitive sampling include (i) estimating the mean or proportion for the most recent occasion, (ii) difference between the means of two consecutive occasions, or (iii) average of means of all the occasions. For measuring changes over two occasions, we generally observe the same sample over two consecutive periods, while fresh samples for each of the occasions are selected for measuring the mean of all the occasions. For estimating the mean or total of the most recent occasion, a portion of the sample selected on the earlier occasions, technically called a matched sample, is retained and it is supplemented by a fresh sample for the most recent occasion.

From the records we find that the pioneering work on sampling over two occasions was done by Jessen (1942) to estimate mean on the most recent occasion. Yates (1949) extended Jessen (1942) study to include cases where sampling is repeated $h(\geq 2)$ occasions. He, however, worked under the restriction of a constant "total sample size" and "matching sampling fraction" over the occasions and postulated the correlation coefficient of the values for the units over k occasions apart to be of the form ρ^k. Some of his restrictions were later relaxed by Patterson (1950) and Tikkiwal (1951). Ecler (1955) stretched the ideas further to develop the theory of rotation sampling. However, in these investigations, the challenge was to estimate a finite population mean, and the optimalities of the suggested sampling strategies were established under the postulation of an infinite population

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00011-X

setup. Kuldoroff (1963) and Rao and Graham (1964) developed a truly finite population theory for estimating the population mean over various occasions. Raj (1965a,b), Ghangurde and Rao (1969), Avadhan and Sukhatme (1970), Chotai (1974), and Chaudhuri and Arnab (1979a), and Arnab (1991, 1998a), among others, contributed theories involving varying sampling schemes over two, as well as more than two, occasions. Besides, Singh (1968), Singh and Kathuria (1969), and Arnab (1980) are among those who developed multistage sampling techniques for two or more occasions. Duncan and Kalton (1987) and Binder and Hidroglou (1988) cited the application of time series analysis in successive sampling for estimating population characteristics at regular intervals of time.

11.2 ESTIMATION OF MEAN FOR THE MOST RECENT OCCASION

Consider a finite population $U = (u_1, \ldots, u_i, \ldots, u_N)$ of N identifiable units, which is supposed to be sampled over various occasions to estimate the population mean for the most recent occasion. Here, we assume that the units of the population are unchanged over various occasions, i.e., the sampling frames remain the same: no new units are added or deleted from the population. Let y_{hi} be the value of the variate under study y for the ith unit of the population on the hth occasion, $Y_h = \sum_{i \in U} y_{hi}$ be the total and $\overline{Y}_h = Y_h/N$ be its mean.

11.2.1 Sampling on Two Occasions

11.2.1.1 Sampling Scheme

On the first occasion a sample s_1 of size n_1 is selected from the population U with probability $p(s_1)$ by some suitable sampling scheme. On the second occasion a subsample (which is technically called the "matched sample") s_{2m} of suitable size $m(\leq n_1)$ is selected from s_1 with probability $p(s_{2m}|s_1)$ by some appropriate sampling scheme, and an unmatched sample s_{2u} of size $u(=n_2 - m)$ is selected with probability $p(s_{2u}|s_1)$ either from the entire population U or from $U - s_1$, the set of units not selected in the first occasions. The sample size m is chosen in an optimal manner viz. considering either the cost of the survey or the efficiency of the estimators. The sample in the second occasion will be denoted by $s_2 = s_{2m} \cup s_{2u}$.

11.2.1.2 General Method of Estimation

Here, we construct two unbiased estimators of $\overline{Y}_1$, the population mean of the first occasion. One of the estimators is based on the initial sample s_1 and the other is based on the matched sample s_{2m}. They are respectively as follows:

$$\widehat{\overline{Y}}_1 = \sum_{i \in s_1} b(s_1, i) y_{1i} \text{ and } t_{1m} = \sum_{i \in s_{2m}} b(s_{2m}, i) y_{1i} \tag{11.2.1}$$

The estimator t_{1m} is such that $E(t_{1m}|s_1) = \widehat{\overline{Y}}_1$, i.e., $b(s_{2m}, i)$'s satisfy $\sum_{s_{2m} \supset i} b(s_{2m}, i) p(s_{2m}|s_1) = b(s_1, i) \, \forall \, i \in s_1$. The unbiasedness conditions of $\widehat{\overline{Y}}_1$ yield $\sum_{s_1 \supset i} b(s_1, i) p(s_1) = 1/N \quad \forall \quad i \in U$.

Similarly, from the data collected on the second occasion through the matched sample s_{2m} and the unmatched sample s_{2u}, two unbiased estimators of the mean $\overline{Y}_2$ of the second occasion are constructed as follows:

$$t_{2m} = \sum_{i \in s_{2m}} b(s_{2m}, i) y_{2i} \text{ and } \widehat{\overline{Y}}_{2u} = \sum_{i \in s_{2u}} b(s_{2u}, i) y_{2i}$$

The constants $b(s_{2u}, i)$'s satisfy the unbiasedness condition $\sum_{s_{2u} \supset i} b(s_{2u}, i) p(s_{2u}|s_1) = 1/N \quad \forall \quad i \in U$. From the estimators $\widehat{\overline{Y}}_1$, t_{1m}, t_{2m}, and $\widehat{\overline{Y}}_{2u}$, a composite estimator of $\overline{Y}_2$ is obtained as

$$\widehat{\overline{Y}}_{2c} = \phi t_{2m} + \psi_1 t_{1m} + \psi_2 \widehat{\overline{Y}}_1 + \varphi \widehat{\overline{Y}}_{2u} \tag{11.2.2}$$

The weights ϕ, ψ_1, ψ_2, and φ are chosen to make $\widehat{\overline{Y}}_{2c}$ unbiased for $\overline{Y}_2$. The condition $E\left(\widehat{\overline{Y}}_{2c}\right) = \overline{Y}_2$ yields

$$\varphi + \phi = 1 \text{ and } \psi_1 + \psi_2 = 0 \tag{11.2.3}$$

On substituting Eq. (11.2.3) in Eq. (11.2.2) we get

$$\widehat{\overline{Y}}_{2c} = \phi \widehat{\overline{Y}}_{2m} + (1 - \phi) \widehat{\overline{Y}}_{2u} \tag{11.2.4}$$

where $\widehat{\overline{Y}}_{2m} = t_{2m} - \beta\left(t_{1m} - \widehat{\overline{Y}}_1\right)$ and $\beta = -\psi_1/\varphi$.

The variance of $\widehat{\overline{Y}}_{2m}$ is

$$V\left(\widehat{\overline{Y}}_{2m}\right) = E\left\{V\left(\widehat{\overline{Y}}_{2m}\middle|s_1\right)\right\} + V\left\{E\left(\widehat{\overline{Y}}_{2m}\middle|s_1\right)\right\}$$

$$= E\{V(t_{2m}|s_1)\} + \beta^2 E\{V(t_{1m}|s_1)\} - 2\beta E\{Cov(t_{2m}, t_{1m}|s_1)\} + V\left\{\sum_{i \in s_1} b(s_1, i) y_{2i}\right\}$$

The optimum value of β that minimizes the variance $V\left(\widehat{\overline{Y}}_{2m}\right)$ is

$$\beta_0 = E\{Cov(t_{1m}, t_{2m}|s_1)\}/E\{V(t_{1m}|s_1)\} \tag{11.2.5}$$

The estimator $\widehat{\overline{Y}}_{2m}$ with $\beta = \beta_0$ will be denoted by

$$\widehat{\overline{Y}}_{2m}^{0} = t_{2m} - \beta_0\left(t_{1m} - \widehat{\overline{Y}}_1\right) \tag{11.2.6}$$

By replacing $\widehat{\overline{Y}}_{2m}$ with $\widehat{\overline{Y}}_{2m}^{0}$ in Eq. (11.2.4), we derive the composite estimator for $\overline{Y}_2$ as

$$\widehat{\overline{Y}}_{2c}^{0} = \phi\widehat{\overline{Y}}_{2m}^{0} + (1-\phi)\widehat{\overline{Y}}_{2u} \tag{11.2.7}$$

Minimizing the variance of $\widehat{\overline{Y}}_{2c}^{0}$ with respect to ϕ, the optimum value of ϕ is obtained as

$$\phi_0 = \left(1/V'_m + 1/V'_u\right)^{-1}\Big/V'_m \tag{11.2.8}$$

where $V'_m = V\left(\widehat{\overline{Y}}_{2m}^{0}\right) - V_{m,u}$, $V'_u = V\left(\widehat{\overline{Y}}_{2u}\right) - V_{m,u}$ and $V_{m,u} = Cov\left(\widehat{\overline{Y}}_{2m}^{0}, \widehat{\overline{Y}}_{2u}\right)$.

Substituting $\phi = \phi_0$ in Eq. (11.2.7), the optimum estimator of $\overline{Y}_2$ is obtained as

$$\widehat{\overline{Y}}_2(opt) = \phi_0\widehat{\overline{Y}}_{2m}^{0} + (1-\phi_0)\widehat{\overline{Y}}_{2u} \tag{11.2.9}$$

The expression for variance of $\widehat{\overline{Y}}_2(opt)$ is obtained as

$$V_{opt} = \left(1/V'_m + 1/V'_u\right)^{-1} + V_{m,u} \tag{11.2.10}$$

In many situations the expression of V_{opt} is obtained as an explicit function of m and u. In this situation we can find the optimum value of m (i.e., $u = n_2 - m$) by minimizing V_{opt} while keeping total sample size for the second occasion fixed or by keeping total cost of the survey to a certain level using a suitable cost function. The optimum values of m and u so obtained will be denoted by m_0 and u_0, respectively. Finally, replacing the optimum value of m_0 and u_0 in the expressions of $\widehat{\overline{Y}}_2(opt)$ and V_{opt}, the final estimator of $\overline{Y}_2$ and its variance are obtained and they are denoted, respectively, by

$$\widehat{\overline{Y}}_2(\min) = \widehat{\overline{Y}}_2(opt)\Big|_{m=m_0, u=u_o} \tag{11.2.11}$$

and

$$V_{\min} = V\left\{\widehat{\overline{Y}}_2(\min)\right\} = V_{opt}\big|_{m=m_0,u=u_o} \tag{11.2.12}$$

The optimum sampling strategy for sampling over two occasions is a combination of the sampling designs used for the selection of samples over two occasions and the estimator $\widehat{\overline{Y}}_2(\min)$ that yields the minimum value of $V_{\min}$.

In the next section we will present a few sampling strategies over two occasions that are commonly used in practice for estimating $\overline{Y}_2$, the mean of the current (second) occasion. The expressions for $V_{\min}$ and the optimum fraction of the matched sample $\lambda_0 = m_0/n_2$ are also derived.

Remark 11.2.1

In most situations the estimators $\widehat{\overline{Y}}_{2m}^{0}$ and $\widehat{\overline{Y}}_2(opt)$ cannot be used in practice because they involve unknown parameters. In practice, we replace unknown parameters with their suitable estimates. For example, we take

$$\widehat{\overline{Y}}_{2m}^{0} = t_{2m} - \widehat{\beta}_0\left(t_{1m} - \widehat{\overline{Y}}_1\right)$$

and

$$\widehat{\overline{Y}}_2(opt) = \widehat{\phi}_0\widehat{\overline{Y}}_{2m}^{0} + \left(1 - \widehat{\phi}_0\right)\widehat{\overline{Y}}_{2u}$$

where $\widehat{\beta}_0 = \widehat{C}ov(t_{1m}, t_{2m}|s_1)\Big/\widehat{V}(t_{1m}|s_1)$; $\widehat{\phi}_0 = \left(1\Big/\widehat{V}'_m + 1\Big/\widehat{V}'_u\right)^{-1}\Big/\widehat{V}'_m$;

and $\widehat{C}ov(t_{1m}, t_{2m}|s_1)$, $\widehat{V}(t_{1m}|s_1)$, $\widehat{V}'_m$, and $\widehat{V}'_u$ are unbiased estimators of $Cov(t_{1m}, t_{2m}|s_1)$, $V(t_{1m}|s_1)$, V'_m, and V'_u, respectively.

11.2.1.3 Simple Random Sampling Without Replacement

On the first occasion, a sample s_1 of size n_1 is selected from the population U by simple random sampling without replacement (SRSWOR) method. On the second occasion, a matched sample s_{2m} of size $m(\le n_1)$ is selected from s_1 by SRSWOR method. The unmatched sample s_{2u} of size $u(=n_2 - m \ge 0)$ is selected from $U - s_1$, the set of $N - n_1$ units that were not selected on the first occasion.

The proposed unbiased estimator for $\overline{Y}_2$ based on the matched sample s_{2m} and the initial sample s_1 is given by

$$\widehat{\overline{Y}}_{2m} = \overline{y}_{2m} - \beta\left(\overline{y}_{1m} - \overline{y}_{1n_1}\right) \tag{11.2.13}$$

where $\overline{y}_{2m}=\sum_{i\in s_{2m}} y_{2i}/m=$ sample mean of y for the second occasion based on s_{2m}, $\overline{y}_{1m}=\sum_{i\in s_{2m}} y_{1i}/m=$ sample mean of y for the first occasion based on s_{2m}, $\overline{y}_{1n_1}=\sum_{i\in s_1} y_{1i}/n_1=$ sample mean of y for the first occasion based on s_1, and β is a suitably chosen constant.

Theorem 11.2.1

(i) $E\left(\widehat{\overline{Y}}_{2m}\right)=\overline{Y}_2$ and (ii) $V\left(\widehat{\overline{Y}}_{2m}\right)=\left(\frac{1}{m}-\frac{1}{n_1}\right)S^2_{d\beta}+\left(\frac{1}{n_1}-\frac{1}{N}\right)S^2_{2y}$

where $S^2_{d\beta}=\sum_{i\in U}\left(d_{i\beta}-\overline{D}_\beta\right)^2/(N-1)$, $d_{i\beta}=y_{2i}-\beta y_{1i}$, $\overline{D}_\beta=\overline{Y}_2-\beta\overline{Y}_1$,

and $S^2_{2y}=\sum_{i\in U}\left(y_{2i}-\overline{Y}_2\right)^2/(N-1)$

Proof

(i) $E\left(\widehat{\overline{Y}}_{2m}\right)=E\left\{E\left(\widehat{\overline{Y}}_{2m}\middle|s_1\right)\right\}=E\left(\overline{y}_{2n_1}\right)=\overline{Y}_2$

$$\left(\text{where } \overline{y}_{2n_1}=\sum_{i\in s_1} y_{2i}/n_1\right)$$

(ii)
$$V\left(\widehat{\overline{Y}}_{2m}\right)=E\left\{V\left(\widehat{\overline{Y}}_{2m}\middle|s_1\right)\right\}+V\left\{E\left(\widehat{\overline{Y}}_{2m}\middle|s_1\right)\right\}$$
$$=\left(\frac{1}{m}-\frac{1}{n_1}\right)E\left(s^2_{d\beta}\right)+V\left(\overline{y}_{2n}\right)$$
$$\left(\text{where } s^2_{d\beta}=\sum_{i\in s_1}\left(d_{i\beta}-\overline{d}_\beta\right)^2\Big/(n_1-1) \text{ and } \overline{d}_\beta=\sum_{i\in s_1} d_{i\beta}/n_1\right)$$
$$=\left(\frac{1}{m}-\frac{1}{n_1}\right)S^2_{d\beta}+\left(\frac{1}{n_2}-\frac{1}{N}\right)S^2_{2y}$$

Now, minimizing $V\left(\widehat{\overline{Y}}_{2m}\right)$ with respect to β, the optimum value of β comes out as

$$\beta=\beta_0=S_{12y}\Big/S^2_{1y} \tag{11.2.14}$$

where $S_{12y}=\sum_{i\in U}\left(y_{1i}-\overline{Y}_1\right)\left(y_{2i}-\overline{Y}_2\right)/(N-1)$ and

$S^2_{1y}=\sum_{i\in U}\left(y_{1i}-\overline{Y}_1\right)^2/(N-1)$.

Placing $\beta = \beta_0$ in Eq. (11.2.13), the optimum estimator $\widehat{\overline{Y}}_{2m}$ and its variances come out as follows:

$$\widehat{\overline{Y}}^{0}_{2m} = \overline{y}_{2m} - \beta_0\left(\overline{y}_{1m} - \overline{y}_{1n_1}\right) \tag{11.2.15}$$

and

$$V\left(\widehat{\overline{Y}}^{0}_{2m}\right) = \left[\left(\frac{1}{m} - \frac{1}{n_1}\right)\rho^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)\right] S^2_{2y} \tag{11.2.16}$$

where $\rho = S_{12y}/(S_{1y}S_{2y})$ = correlation coefficient between y_1 and y_2.

An unbiased estimator for $\overline{Y}_2$ based on the unmatched sample s_{2u} is given by

$$\widehat{\overline{Y}}_{2u} = \overline{y}_{2u} = \sum_{i \in s_{2u}} y_{2i}/u \tag{11.2.17}$$

where $u = n_2 - m$.

The composite estimator for $\overline{Y}_2$ is given by

$$\begin{aligned}\widehat{\overline{Y}}_{2c} &= \phi\widehat{\overline{Y}}^{0}_{2m} + (1-\phi)\widehat{\overline{Y}}_{2u} \\ &= \phi\left\{\overline{y}_{2m} - \beta_0\left(\overline{y}_{1m} - \overline{y}_{1n_1}\right)\right\} + (1-\phi)\overline{y}_{2u}\end{aligned} \tag{11.2.18}$$

Theorem 11.2.2

The optimum value ϕ that minimizes $V\left(\widehat{\overline{Y}}_{2c}\right)$ and the corresponding value of $V\left(\widehat{\overline{Y}}_{2c}\right)$ are, respectively,

$$\phi_0 = \left\{\left(\frac{1}{m} - \frac{1}{n_1}\right)(1-\rho^2) + \frac{1}{n_1}\right\}^{-1}\left[\left\{\left(\frac{1}{m} - \frac{1}{n_1}\right)(1-\rho^2) + \frac{1}{n_1}\right\}^{-1} + u\right]^{-1}$$

and

$$V_{opt} = \left[\left\{\left(\frac{1}{m} - \frac{1}{n_1}\right)(1-\rho^2) + \frac{1}{n_1}\right\}^{-1} + u\right]^{-1} S^2_{2y} - \frac{S^2_{2y}}{N}$$

Proof

$$
\begin{aligned}
V\left(\widehat{\overline{Y}}_{2m}^{0}\right) &= E\left[V\left\{\overline{y}_{2m} - \beta_0\left(\overline{y}_{1m} - \overline{y}_{1n_1}\right)\Big|s_1\right\}\right] \\
&\quad + V\left[E\left\{\overline{y}_{2m} - \beta_0\left(\overline{y}_{1m} - \overline{y}_{1n_1}\right)\Big|s_1\right\}\right] \\
&= \left(\frac{1}{m} - \frac{1}{n_1}\right)(1-\rho^2)S_{2y}^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)S_{2y}^2
\end{aligned} \tag{11.2.19}
$$

$$
\begin{aligned}
V\left(\overline{y}_{2u}\right) &= E\left\{V\left(\overline{y}_{2u}\big|s_1\right)\right\} + V\left\{E\left(\overline{y}_{2u}\big|s_1\right)\right\} \\
&= \left(\frac{1}{u} - \frac{1}{N-u}\right)E\left[\frac{1}{N-u-1}\sum_{i\in U-s_1}\left(y_{2i} - \sum_{i\in U-s_1} y_{2i}/(N-u)\right)^2\right] \\
&\quad + V\left(\sum_{i\in U-s_1} y_{2i}/(N-u)\right) \\
&= \left(\frac{1}{u} - \frac{1}{N}\right)S_{2y}^2
\end{aligned} \tag{11.2.20}
$$

and

$$
\begin{aligned}
Cov\left(\widehat{\overline{Y}}_{2m}^{0}, \widehat{\overline{Y}}_{2u}\right) &= Cov\left(\overline{y}_{2m}, \overline{y}_{2u}\right) \\
&= E\left\{Cov\left(\overline{y}_{2m}, \overline{y}_{2u}\big|s_1\right)\right\} \\
&\quad + Cov\left\{E\left(\overline{y}_{2m}\big|s_1\right), E\left(\overline{y}_{2u}\big|s_1\right)\right\} \\
&= Cov\left(\sum_{i\in s_1} y_{2i}/n_1, \sum_{i\in U-s_1} y_{2i}/(N-n_1)\right) \\
&= -n_1^2 V\left(\overline{y}_{2n_1}\right)\Big/\left\{n_1(N-n_1)\right\} \\
&\quad \left(\text{where } \overline{y}_{2n_1} = \sum_{i\in s_1} y_{2i}/n_1\right) \\
&= -S_{2y}^2\Big/N = V_{m,u}
\end{aligned} \tag{11.2.21}
$$

Writing $V'_m = V\left(\widehat{\overline{Y}}{}^0_{2m}\right) - V_{m,u} = \left[\left(\frac{1}{m} - \frac{1}{n_1}\right)(1-\rho^2) + \frac{1}{n_1}\right] S^2_{2y}$, $V'_u = V\left(\widehat{\overline{Y}}_{2u}\right) - V_{m,u} = S^2_{2y}\big/u$ and using Eqs. (11.2.8) and (11.2.10), we get

$$\phi_0 = \left(1/V'_m + 1/V'_u\right)^{-1}\Big/V'_m$$

$$= \left\{\left(\frac{1}{m} - \frac{1}{n_1}\right)(1-\rho^2) + \frac{1}{n_1}\right\}^{-1}\left[\left\{\left(\frac{1}{m} - \frac{1}{n_1}\right)(1-\rho^2) + \frac{1}{n_1}\right\}^{-1} + u\right]^{-1} \tag{11.2.22}$$

and

$$V_{opt} = \left(1/V'_m + 1/V'_u\right)^{-1} + V_{m,u}$$

$$= \left[\left\{\left(\frac{1}{m} - \frac{1}{n_1}\right)(1-\rho^2) + \frac{1}{n_1}\right\}^{-1} + u\right]^{-1} S^2_{2y} - S^2_{2y}\big/N \tag{11.2.23}$$

11.2.1.3.1 Optimum Allocation of the Matched Sample

Theorem 11.2.3

For fixed n_1 and n_2, the optimum value of the proportion of the matched sample $\lambda = m/n_2$ and the corresponding value of V_{opt} are given, respectively, by

$$\lambda_0 = \gamma\sqrt{1-\rho^2}\left(1+\sqrt{1-\rho^2}\right)^{-1} \text{ and}$$

$$V_{\min} = \left[\frac{1}{n_2}\left\{\gamma\frac{2}{1+\sqrt{1-\rho^2}} + (1-\gamma)\right\}^{-1} - \frac{1}{N}\right] S^2_{2y}$$

where $\gamma = n_1/n_2$.

Proof

Putting $u = n_2 - m$ in the expression of V_{opt} in Eq. (11.2.23) and then differentiating it with respect to m and equating the result to zero, we get

$$\frac{(1-\rho^2)}{m^2}\left[\left(\frac{1}{m}-\frac{1}{n_1}\right)(1-\rho^2)+\frac{1}{n_1}\right]^{-2} = 1 \tag{11.2.24}$$

Eq. (11.2.24) yields the optimum value of λ as
$\lambda_0 = \gamma\sqrt{1-\rho^2}\left(1+\sqrt{1-\rho^2}\right)^{-1}$.

Finally, placing $\lambda = \lambda_0$ in the expression of V_{opt}, we obtain

$$V_{\min} = \left[\frac{1}{n_2}\left\{\gamma\frac{2}{1+\sqrt{1-\rho^2}}+(1-\gamma)\right\}^{-1}-\frac{1}{N}\right]S_{2y}^2$$

Remark 11.2.2

The estimators $\widehat{\overline{Y}}{}_{2m}^{0}$, $\widehat{\overline{Y}}_2(opt)$, and the optimum matching fraction λ_0 cannot be used in practice because they involve unknown parameters S_{1y}^2, S_{2y}^2, β_0, and ρ. So, in all practical purposes we replace the parameters by their respective estimates $s_{1y}^2 = \sum_{i\in s_{2m}}\left(y_{1i}-\overline{y}_{1m}\right)^2/(m-1)$, $s_{2y}^2 = \sum_{i\in s_{2m}\cup s_{2u}}\left(y_{2i}-\overline{y}_2\right)^2/(n_2-1)$, $\widehat{\beta}_0 = s_{12y}/s_{1y}^2$, and $\widehat{\rho} = s_{12y}/(s_{1y}\,s_{2y})$, where $s_{12y} = \sum_{i\in s_{2m}}\left(y_{1i}-\overline{y}_{1m}\right)\left(y_{2i}-\overline{y}_{2m}\right)/(m-1)$ and $\overline{y}_2 = \sum_{i\in s_{2m}\cup s_{2u}} y_{2i}/n_2$.

Remark 11.2.3

If no information about the first occasion is used for estimating the second occasion, i.e., in the second occasion an unmatched sample s_2 of size n_2 is selected, then we would use $\overline{y}(s_2) = \sum_{i\in s_2} y_{2i}/n_2$ as an estimator of the population mean $\overline{Y}_2$. The efficiency of the optimum estimator based on the partially matched sample compared to the totally unmatched sample of size n_2 is given by

$$E = \frac{V\{\overline{y}(s_2)\}}{V_{\min}} = (1-f_2)\left[\left\{\gamma\frac{2}{1+\sqrt{1-\rho^2}}+(1-\gamma)\right\}^{-1}-f_2\right]^{-1} \tag{11.2.25}$$

where $f_2 = n_2/N$.

The efficiency $E \geq 1$ as $\rho^2 \leq 1$ and it increases with ρ^2. Hence partial matching produces a gain in efficiency unless $\rho = 0$. The efficiency E attains a maximum value $1 + \dfrac{\gamma}{1-(1+\gamma)f_2}$ when $|\rho| = 1$.

Remark 11.2.4

The optimum proportion of matched sample λ_0 decreases as $|\rho|$ increases. λ_0 attains a maximum $\gamma/2$ when $|\rho| = 0$ and the minimum value of λ_0 is zero when is $|\rho| = 1$.

Remark 11.2.5

The minimum variance $V_{\min}$ decreases with increase of $|\rho|$ and it attains a minimum value $\left(\dfrac{1}{n_2(1+\gamma)} - \dfrac{1}{N}\right) S_{2y}^2$ when the correlation ρ is either $+1$ or -1, and in this case, $\lambda_0 = 0$. But when $\lambda_0 = 0$, no matched sample is used and in this case $\widehat{\overline{Y}}_{2c} = \overline{y}(s_2)$ and $V_{\min} = (1 - f_2)S_{2y}^2 / n_2$. In fact, the expression of $V_{\min}$ was derived assuming m, $u > 1$. So, when $|\rho|$ is very high we should take the matched sample m to be as small as possible with a minimum of 2 to estimate a parameter like β_0. Similarly, when ρ is very small, we should take m to be as large as possible and a small unmatched sample with a minimum of 2.

Remark 11.2.6

In case $n_1 = n_2 = n$, the expression for the minimum variance is obtained as

$$V_{\min} = \left(\frac{1+\sqrt{1-\rho^2}}{2n} - \frac{1}{N}\right) S_{2y}^2 \tag{11.2.26}$$

Remark 11.2.7

If $n_1 = n_2 = n$ and the unmatched sample s_{2u} is selected from the entire population, then s_1 and s_{2u} become independent and we have $V_{m,u} = 0$. In this situation

$$V_{\min} = \frac{1+\sqrt{1-\rho^2}}{2n} S_{2y}^2 \tag{11.2.27}$$

Remark 11.2.8

Because the magnitude of Eq. (11.2.27) is larger than that of Eq. (11.2.26), we conclude that the selection of the unmatched sample from those units that are not selected in the first occasion produces a gain in efficiency of the combined estimator.

11.2.1.4 Probability Proportional to Size With Replacement Sampling

Raj (1965a,b) proposed the following sampling scheme over two occasions: On the first occasion a sample s_1 of size n is selected from the population U using probability proportional to size with replacement (PPSWR) sampling with known measure of size $z_i(>0)$ for the ith unit, $i = 1,\ldots, N$. On the second occasion a matched subsample s_{2m} of size $m(\leq n)$ is selected from s_1 by SRSWOR method, treating all the selected units in s_1 as distinct. The unmatched sample s_{2u} of size $u(=n-m)$ is selected independently from the entire population by PPSWR method using z_i as the size of measure for the ith unit.

Chaudhuri and Arnab (1979a) proposed the following estimator for Y_2, which is more efficient than the Raj (1965a,b) estimator:

$$\widehat{\widehat{Y}}_{2c} = \phi \widehat{Y}_{2m} + (1-\phi)\widehat{Y}_{2u} \tag{11.2.28}$$

where $\widehat{Y}_{2m} = \frac{1}{m}\sum_{i\in s_{2m}}\frac{y_{2i}}{p_i} - \beta\left(\frac{1}{m}\sum_{i\in s_{2m}}\frac{y_{1i}}{p_i} - \frac{1}{n}\sum_{i\in s_1}\frac{y_{1i}}{p_i}\right)$, $\widehat{Y}_{2u} = \frac{1}{u}\sum_{i\in s_{2u}}\frac{y_{2i}}{p_i}$, $p_i = \frac{z_i}{Z}$, $Z = \sum_{i\in U} z_i$, $\sum_{i\in s_t}$ denotes the sum over the units in s_t with repetition, and β is a constant chosen to minimize $V\left(\widehat{Y}_{2m}\right)$.

Theorem 11.2.4

(i) $E\left(\widehat{Y}_{2m}\right) = Y_2$

(ii) $V\left(\widehat{Y}_{2m}\right) = \left(\frac{1}{m} - \frac{1}{n}\right)\left(V_{2|z} - 2\beta\delta\sqrt{V_{1|z}\cdot V_{2|z}} + \beta^2 V_{1|z}\right) + \frac{1}{n}V_{2|z}$

(iii) The optimum value of β that minimizes $V\left(\widehat{Y}_{2m}\right)$ is $\beta_0 = \delta\sqrt{V_{2|z}/V_{1|z}}$

(iv) $V\left(\widehat{Y}^0_{2m}\right) = \left(\frac{1}{m} - \frac{1}{n}\right)\left(1-\delta^2\right)V_{2|z} + \frac{1}{n}V_{2|z}$

where $\widehat{Y}^0_{2m} = \frac{1}{m}\sum_{i\in s_{2m}}\frac{y_{2i}}{p_i} - \beta_0\left(\frac{1}{m}\sum_{i\in s_{2m}}\frac{y_{1i}}{p_i} - \frac{1}{n}\sum_{i\in s_1}\frac{y_{1i}}{p_i}\right)$,

$V_{1|z} = \sum_{i\in U} p_i\left(\frac{y_{1i}}{p_i} - Y_1\right)^2$, $\quad V_{2|z} = \sum_{i\in U} p_i\left(\frac{y_{2i}}{p_i} - Y_2\right)^2$, and

$\delta = \sum_{i\in U} p_i\left(\frac{y_{1i}}{p_i} - Y_1\right)\left(\frac{y_{2i}}{p_i} - Y_2\right)\Big/\sqrt{\left(V_{1|z}\cdot V_{2|z}\right)}$.

Proof

(i) $E\left(\widehat{Y}_{2m}\right) = E\left\{E\left(\widehat{Y}_{2m}\middle|s_1\right)\right\} = E\left(\frac{1}{n}\sum_{i\in s_1}\frac{y_{2i}}{p_i}\right) = Y_2$

(ii) $V\left(\widehat{Y}_{2m}\right) = E\left\{V\left(\widehat{Y}_{2m}\middle|s_1\right)\right\} + V\left\{E\left(\widehat{Y}_{2m}\middle|s_1\right)\right\}$

$$= \left(\frac{1}{m}-\frac{1}{n}\right)\frac{1}{n-1}E\left[\sum_{i\in s_1}\left\{\frac{y_{2i}-\beta y_{1i}}{p_i}-\frac{1}{n}\sum_{i\in s_1}\frac{y_{2i}-\beta y_{1i}}{p_i}\right\}^2\right]$$

$$+ V\left(\frac{1}{n}\sum_{i\in s_1}\frac{y_{2i}}{p_i}\right)$$

$$= \left(\frac{1}{m}-\frac{1}{n}\right)\sum_{i\in U}\left\{\frac{y_{2i}-\beta y_{1i}}{p_i}-\left(Y_2-\beta Y_1\right)\right\}^2 + V\left(\frac{1}{n}\sum_{i\in s_1}\frac{y_{2i}}{p_i}\right)$$

$$= \left(\frac{1}{m}-\frac{1}{n}\right)\left(V_{2|z}-2\beta\delta\sqrt{V_{1|z}\cdot V_{2|z}}+\beta^2 V_{1|z}\right)+\frac{1}{n}V_{2|z}$$

(iii) Differentiating $V\left(\widehat{Y}_{2m}\right)$ with respect to β and equating it to zero we get the optimum value of $\beta = \beta_0 = \delta\sqrt{V_{2|z}/V_{1|z}}$

(iv) Putting $\beta = \beta_0$ in the expressions of $\widehat{Y}_{2m}$ and $V\left(\widehat{Y}_{2m}\right)$, we get

$$\widehat{Y}^0_{2m} = \frac{1}{m}\sum_{i\in s_{2m}}\frac{y_{2i}}{p_i}-\beta_0\left(\frac{1}{m}\sum_{i\in s_{2m}}\frac{y_{1i}}{p_i}-\frac{1}{n}\sum_{i\in s_1}\frac{y_{1i}}{p_i}\right)$$

and

$$V\left(\widehat{Y}^0_{2m}\right) = \left(\frac{1}{m}-\frac{1}{n}\right)\left(1-\delta^2\right)V_{2|z}+\frac{1}{n}V_{2|z}$$

Theorem 11.2.5

The optimum proportion of the matched sample $\lambda = m/n$ that minimizes the variance of $\widehat{\widehat{Y}}_{2c}$ and the expression of minimum variance of $\widehat{\widehat{Y}}_{2c}$ are, respectively,

(i) $\lambda_0 = \dfrac{\sqrt{1-\delta^2}}{1+\sqrt{1-\delta^2}}$ and (ii) $V_{\min} = \dfrac{1+\sqrt{1-\delta^2}}{2n}V_{2|z}$

Proof

Putting $\widehat{Y}_{2m}=\widehat{Y}^{0}_{2m}$ in the expression of $\widehat{\widehat{Y}}_{2c}$ and then minimizing the variance of $\widehat{Y}^{0}_{2c}=\phi\widehat{Y}^{0}_{2m}+(1-\phi)\widehat{Y}_{2u}$ with respect to ϕ, the optimum value of ϕ is obtained as

$$\phi=\phi_0=\frac{1/V\left(\widehat{Y}^{0}_{2m}\right)}{1/V\left(\widehat{Y}^{0}_{2m}\right)+1/V\left(\widehat{Y}^{0}_{2u}\right)}$$

Now, noting $V\left(\widehat{Y}_{2u}\right)=V_{2|z}/u$, the optimum value of $V\left(\widehat{Y}^{0}_{2c}\right)$ with $\phi=\phi_0$ is obtained as

$$V_{opt}=\left[\left\{\left(\frac{1}{m}-\frac{1}{n}\right)\left(1-\delta^2\right)+\frac{1}{n}\right\}^{-1}+u\right]^{-1}V_{2|z} \qquad (11.2.29)$$

Minimizing V_{opt} with respect to m while keeping $m+u=n$ as fixed, the optimum proportion of the matched sample is obtained as

$$\lambda_0=\sqrt{1-\delta^2}\left(1+\sqrt{1-\delta^2}\right)^{-1}$$

Finally, substituting $m=n\lambda_0$ in Eq. (11.2.29), the expression for the minimum variance of $\widehat{\widehat{Y}}_{2c}$ comes out as

$$V_{\min}=\frac{1+\sqrt{1-\delta^2}}{2n}V_{2|z}$$

Remark 11.2.9

Here, we note that the optimum proportion of matched sample λ_0 always lies between zero and one. In case information on the first occasion is not used and an unmatched sample s_2 of size n is selected by PPSWR method, one would use an unbiased estimator of the total Y_2 as $\widehat{Y}_{2n}=\frac{1}{n}\sum_{i\in S_2}\frac{y_{2i}}{p_i}$. Let $\widehat{Y}^{0}_{2c}$ be $\widehat{Y}_{2c}$ with the optimum values β, m, and ϕ. Then the relative efficiency of $\widehat{Y}^{0}_{2c}$ with respect to $\widehat{Y}_{2n}$ is given by

$$E=V\left(\widehat{Y}_{2n}\right)/V_{\min}=\frac{2}{1+\sqrt{1-\delta^2}}$$

The estimator $\widehat{Y}^{0}_{2c}$ remains more efficient than $\widehat{Y}_{2n}$ for all values of δ. The efficiency attains a maximum 2 when $\delta=\pm1$ and a minimum 1 for $\delta=0$.

11.2.1.5 Simple Random Sampling With Replacement

Let the initial sample s_1 of size n be selected by simple random sampling with replacement (SRSWR) method and the matched sample s_{2m} of size $m(\leq n)$ be selected from s_1 by SRSWOR method, treating all the units in s_1 as distinct, the unmatched sample s_{2u} of size u is selected from the entire population by SRSWR method. Because PPSWR reduces to SRSWR when $p_i = 1/N$, we derive Theorems 11.2.6 and 11.2.7 by substituting $p_i = 1/N$ in Theorems 11.2.4 and 11.2.5 as follows:

Theorem 11.2.6

(i) $\widehat{Y}_{2m} = N\left[\overline{y}_{2m} - \beta\left(\overline{y}_{1m} - \overline{y}_{1n}\right)\right]$ is unbiased for Y_2

(ii) $V\left(\widehat{Y}_{2m}\right) = N^2\left[\left(\frac{1}{m} - \frac{1}{n}\right)\left(\sigma_{2y}^2 - 2\beta\rho\sigma_{1y}\sigma_{2y} + \beta^2\sigma_{1y}^2\right) + \frac{1}{n}\sigma_{2y}^2\right]$

(iii) The optimum value of β that minimizes $V\left(\widehat{Y}_{2m}\right)$ is $\beta_0 = \rho\sigma_{2y}/\sigma_{1y}$

(iv) The variance of $\widehat{Y}_{2m}^0$ is $V\left(\widehat{Y}_{2m}^0\right) = N^2\left[\left(\frac{1}{m} - \frac{1}{n}\right)(1-\rho^2) + \frac{1}{n}\right]\sigma_{2y}^2$

where $\overline{y}_{1n} = \sum_{i\in s_1} y_{1i}/n$, $\overline{y}_{1m} = \sum_{i\in s_{2m}} y_{1i}/m$, $\overline{y}_{2m} = \sum_{i\in s_{2m}} y_{2i}/m$,
$\sigma_{1y}^2 = \sum_{i\in U}\left(y_{1i} - \overline{Y}_1\right)^2/N$, $\sigma_{2y}^2 = \sum_{i\in U}\left(y_{2i} - \overline{Y}_2\right)^2/N$,
$\widehat{Y}_{2m}^0 = N\left\{\overline{y}_{2m} - \beta_0\left(\overline{y}_{1m} - \overline{y}_{1n}\right)\right\}$, and ρ is the correlation coefficient between x and y.

Theorem 11.2.7

(i) $\widehat{Y}_{2c} = \phi\widehat{Y}_{2m}^0 + (1-\phi)\widehat{Y}_{2u}$ is unbiased for Y_2

(ii) The optimum proportion of the matched sample $\lambda_0 = \dfrac{\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}}$

(iii) The minimum variance of $\widehat{Y}_{2c}$ is

$$V_{\min} = N^2\left(\frac{1+\sqrt{1-\rho^2}}{2n}\right)\sigma_{2y}^2$$

where $\widehat{Y}_{2u} = N\overline{y}_{2u}$ and $\overline{y}_{2u} = \sum_{i\in s_{2u}} y_{2i}/u$.

11.2.1.6 Sampling Over Two Occasions: Stratifying the Initial Sample

Here, we suppose that the population should be stratified into homogeneous strata according to certain criteria, but it cannot be employed as we have no information regarding which unit falls in which particular stratum. For instance, in household surveys, households can be stratified effectively according to various income groups viz. high, upper middle, lower middle, and low, but the income levels of the households may not be known before conducting the survey. Hence for estimation of population characteristics and selection of sample in the second occasion, we may use y_{1i}'s for $i \in s_1$, as a stratification variable.

Let, on the first occasion, a PPSWR sample s_1 of size n be selected from a population U using known size measure z_i's. Using the observed values of y_{1i}'s for $i \in s_1$, the selected n sampled units are assigned to one of the L strata $H_1,\ldots, H_l,\ldots, H_L$. Let y_{1j}^l, y_{2j}^l, and z_j^l be the value of the jth unit of the lth stratum for the variables y_1, y_2 and z, respectively. Typically a random number $n_l\left(0 \le n_l \le n, \sum_{l=1}^{L} n_l = n\right)$ of these n units will constitute a sample from the lth strum s_1^l (say). On the second occasion, independent subsamples s_{2m}^l's of sizes m_l's $= m\, n_l/n$ (assumed integer) are chosen from s_1^l $(l = 1,\ldots, L)$ by SRSWOR method. Here, n is assumed to be so large that $Prob(n_l \ge 1) = 1$. Finally, an unmatched sample s_{2u} of size $u = (n - m)$ is selected by PPSWR method from the entire population U using z_i's as a measure of size.

The proposed composite estimator for Y_2 is

$$\widehat{Y}_{2c}(st) = \phi \widehat{Y}_{2m}(st) + (1-\phi)\widehat{Y}_{2u} \tag{11.2.30}$$

where $\widehat{Y}_{2m}(st) = \sum_{l=1}^{L} w_l \widehat{Y}_{2m}^l$, $\quad \widehat{Y}_{2m}^l = \frac{1}{m_l}\sum_{j\in s_{2m}^l}\frac{d_{\beta j}^l}{p_j^l} + \beta^l \frac{1}{n_l}\sum_{j\in s_1^l}\frac{y_{1j}^l}{p_j^l}$,

$\widehat{Y}_{2u} = \frac{1}{u}\sum_{i\in S_{2u}}\frac{y_{2i}}{p_i}$, $\ w_l = n_l/n$, $\ d_{\beta j}^l = y_{2j}^l - \beta^l y_{1j}^l$, $\ p_j^l = z_j^l\big/Z$, $\ p_i = z_i/Z$, $Z = \sum_{j\in U} z_j$, and β^l is a suitably chosen constant.

Theorem 11.2.8

(i) $E\left\{\widehat{Y}_{2m}(st)\right\} = Y_2$

(ii) $V\left\{\widehat{Y}_{2m}(st)\right\} = \left(\frac{1}{m} - \frac{1}{n}\right)\sum_{l=1}^{L}\frac{V_{d|z}^l}{P^l} + \frac{V_{y_2|z}}{n}$

(iii) The optimum value of β^l that minimizes $V\left\{\widehat{Y}_{2m}^l\right\}$ is

$$\beta^{l0} = \delta^l = \left(\sum_{j\in H_l} P_l\frac{y_{1j}y_{2j}}{p_j^l} - Y_1^l Y_2^l\right)\Bigg/\left(V_{y_1|z}^l V_{y_2|z}^l\right)^{1/2}$$

where $V_{d|z}^l = \sum_{j\in H_l} P_l\frac{\left(d_{\beta j}^l\right)^2}{p_j^l} - \left(D_\beta^l\right)^2$, $D_\beta^l = \sum_{j\in H_l} d_{\beta j}^l$,

$V_{y_1|z}^l = \sum_{j\in H_l} P_l\frac{\left(y_{1j}^l\right)^2}{p_j^l} - \left(Y_1^l\right)^2$, $V_{y_2|z}^l = \sum_{j\in H_l} P_l\frac{\left(y_{2j}^l\right)^2}{p_j^l} - \left(Y_2^l\right)^2$, and $P_l = \sum_{j\in H_l} p_j^l$.

Proof

Let $E_1(V_1)$ be the expectation (variance) over the random vector $\mathbf{n} = (n_1, \ldots, n_l)$ while $E_2(V_2)$ and $E_3(V_3)$ denote the conditional expectations (variances) over s_1^l for fixed $\mathbf{n}$ and over s_{2m}^l for fixed s_1^l and $\mathbf{n}$, respectively.

$$
\begin{aligned}
\text{(i)} \qquad E\left\{\widehat{Y}_{2m}(st)\right\} &= E_1\left[\sum_{l=1}^{L} w_l E_2\left\{E_3\left(\widehat{Y}^{l}_{2m}\middle| s_1^l, \mathbf{n}\right)\middle|\mathbf{n}\right\}\right] \\
&= E_1\left[\sum_{l=1}^{L} w_l E_2\left(\frac{1}{n_l}\sum_{j\in S_1^l}\frac{y_{2j}^l}{p_j^l}\middle|\mathbf{n}\right)\right] \\
&= E_1\left(\sum_{l=1}^{L} w_l \frac{Y_2^l}{P_l}\right) \qquad\qquad (11.2.31)\\
&= \sum_{l=1}^{L} Y_2^l \\
&= Y_2
\end{aligned}
$$

(noting $Y_2^l = \sum_{j\in H_l} y_{2j}^l$ and $E(w_l) = P_l$)

$$
\begin{aligned}
\text{(ii)} \qquad V\left\{\widehat{Y}_{2m}(st)\right\} &= E_1\left[V_2\left\{E_3\sum_{l=1}^{L} w_l\left(\widehat{Y}^{l}_{2m}\right)\middle| s_1^l, \mathbf{n}\right\}\right] \\
&\quad + E_1\left[E_2\left\{V_3\sum_{l=1}^{L} w_l\left(\widehat{Y}^{l}_{2m}\right)\middle| s_1^l, \mathbf{n}\right\}\right] \qquad\qquad (11.2.32)\\
&\quad + V_1\left[E_2\left\{\sum_{l=1}^{L} w_l E_3\left(\widehat{Y}^{l}_{2m}\right)\middle| s_1^l, \mathbf{n}\right\}\right]
\end{aligned}
$$

Now,

$$E_1\left[V_2\left\{E_3\sum_{l=1}^{L}w_l\left(\widehat{Y}^l_{2m}\right)\Big|s_1^l,\mathbf{n}\right\}\right]=E_1\left[V_2\left\{\sum_{l=1}^{L}w_l\left(\frac{1}{n_l}\sum\nolimits_{j\in s_1^l}\frac{y_{2j}}{p_j^l}\right)\Bigg|\mathbf{n}\right\}\right]$$

$$=E_1\left(\sum_{l=1}^{L}w_l^2\frac{V^l_{y_2|z}}{n^lP_l^2}\right)$$

$$=\frac{1}{n}E_1\left(\sum_{l=1}^{L}w_l\frac{V^l_{y_2|z}}{P_l^2}\right)$$

$$=\frac{1}{n}\sum_{l=1}^{L}\frac{V^l_{y_2|z}}{P_l}\tag{11.2.33}$$

and

$$E_1\left[E_2\left\{V_3\sum_{l=1}^{L}w_l\left(\widehat{Y}^l_{2m}\right)\Big|s_1^l,\mathbf{n}\right\}\right]=E_1\left[E_2\left\{\sum_{l=1}^{L}w_l^2\left(\frac{1}{m_l}-\frac{1}{n_l}\right)\ \frac{1}{n_l-1}\right.\right.$$

$$\left.\left.\sum_{j\in s_1^l}\left(\frac{d^l_{\beta j}}{p_j^l}-\overline{d}^l_\beta\right)^2\Bigg|\mathbf{n}\right\}\right]$$

$$\left(\text{where }\ \overline{d}^l_\beta=\sum_{j\in s_1^l}\frac{d^l_{\beta j}}{p_j^l}\Big/n_l\right)$$

Now, noting $m_l=mn_l/n$, we get

$$E_1\left[E_2\left\{V_3\sum_{l=1}^{L}w_l\left(\widehat{Y}^l_{2m}\right)\Big|s_1^l,\mathbf{n}\right\}\right]=\left(\frac{1}{m}-\frac{1}{n}\right)E_1\left(\sum_{l=1}^{L}w_l\frac{V^l_{d|z}}{P_l^2}\right)$$

$$=\left(\frac{1}{m}-\frac{1}{n}\right)\sum_{l=1}^{L}\frac{V^l_{d|z}}{P_l}.\tag{11.2.34}$$

Furthermore,

$$V_1\left[E_2\left\{\sum_{l=1}^{L} w_l E_3\left(\widehat{Y}^l_{2m}\right)\middle| s^l_1, \mathbf{n}\right\}\right] = V_1\left[E_2\left\{\sum_{l=1}^{L} w_l\left(\frac{1}{n_l}\sum\nolimits_{j\in s^l_1}\frac{y_{2j}}{p^l_j}\right)\middle|\mathbf{n}\right\}\right]$$

$$= V_1\left(\sum_{l=1}^{L} w_l\frac{Y^l_2}{P_l}\right)$$

$$= \frac{1}{n}\left(\sum_{l=1}^{L}\frac{1-P_l}{P_l}\left(Y^l_2\right)^2 - \sum_{l\neq}^{L}\sum_{k=1}^{L} Y^l_2 Y^k_2\right) \tag{11.2.35}$$

(noting $V(w_l) = P_l(1-P_l)$ and $Cov(w_l, w_{l'}) = -P_l P_{l'}$)
Substituting Eqs. (11.2.33)−(11.2.35) in Eq. (11.2.32), we get

$$V\left\{\widehat{Y}_{2m}(st)\right\} = \frac{1}{n}\sum_{l=1}^{L}\frac{V^l_{y_2|z}}{P_l} + \left(\frac{1}{m}-\frac{1}{n}\right)\sum_{l=1}^{L}\frac{V^l_{d|z}}{P_l}$$
$$+\frac{1}{n}\left(\sum_{l=1}^{L}\frac{1-P_l}{P_l}\left(Y^l_2\right)^2 - \sum_{l\neq}^{L}\sum_{k=1}^{L} Y^l_2 Y^k_2\right)$$

Furthermore, noting $\dfrac{V^l_{y_2|z}}{P_l} = \sum\limits_{j\in H_l}\dfrac{\left(y^l_{2j}\right)^2}{p^l_j} - \dfrac{\left(Y^l_2\right)^2}{P_l}$, we get

$$V\left\{\widehat{Y}_{2m}(st)\right\} = \left(\frac{1}{m}-\frac{1}{n}\right)\sum_{l=1}^{L}\frac{V^l_{d|z}}{P_l} + \frac{1}{n}V_{y_2|z}$$

(iii) Minimizing $V\left\{\widehat{Y}_{2m}(st)\right\}$ with respect to β^l, we arrive at the optimum value of β^l as

$$\beta^{l0} = \delta^l = \left(\sum_{j\in H_l} P^l\frac{y_{1j}y_{2j}}{p^l_j} - Y^l_1 Y^l_2\right)\Big/\left(V^l_{y_2|z}V^l_{y_1|z}\right)^{1/2}$$

Theorem 11.2.9

The optimum proportion of the matched sample and the expression for the minimum variance of $\widehat{Y}_{2c}(st)$ are given, respectively, by

$$opt\ \lambda = \lambda_0 = \frac{\sqrt{A}}{1+\sqrt{A}} \quad\text{and}\quad V_{\min}\left\{\widehat{Y}_{2c}(st)\right\} = \frac{1+\sqrt{A}}{2\,n}V_{y_2|z}$$

where $A = \sum\limits_{l=1}^{L}\dfrac{\left(1-\delta^2_l\right)V^l_{y_2|z}}{V_{y_2|z}P_l}$.

Proof

The estimator $\widehat{Y}_{2m}(st)$ with $\beta^l = \beta^{l0}$ and its variance are given, respectively, by

$$\widehat{Y}^0_{2m}(st) = \sum_{l=1}^{L} w_l \left(\frac{1}{m_l} \sum\nolimits_{j \in s^l_{2m}} \frac{y^l_{2j} - \delta^l y^l_{1j}}{p^l_j} + \delta^l \frac{1}{n_l} \sum\nolimits_{j \in s^l_1} \frac{y^l_{1j}}{p^l_j} \right)$$

and

$$V\left(\widehat{Y}^0_{2m}\right) = \frac{n-m}{mn} \sum_{l=1}^{L} \frac{\left(1-\delta^2_l\right) V^l_{y_2|z}}{P_l} + \frac{V_{y_2|z}}{n}$$

$$= \frac{1}{n}\left(\frac{n-m}{m} A + 1\right) V_{y_2|z}$$

Substituting $\widehat{Y}_{2m}(st) = \widehat{Y}^0_{2m}(st)$ in Eq. (11.2.30) and then minimizing the variance $\widehat{Y}_{2c}(st)$ with respect to ϕ, the optimum variance $\widehat{Y}_{2c}(st)$ comes out as

$$V_{opt}\left\{\widehat{Y}_{2c}(st)\right\} = \left[n \left\{ \frac{n-m}{m} A + 1 \right\}^{-1} + u \right]^{-1} V_{y_2|z} \tag{11.2.36}$$

Now, minimizing $V_{opt}\left\{\widehat{Y}_{2c}(st)\right\}$ with respect to m and keeping n as fixed, the optimum proportion of matched sample $\lambda = m/n$ and the corresponding value of $V_{opt}\left\{\widehat{Y}_{2c}(st)\right\}$ are obtained, respectively, as

$$\lambda = \lambda_0 = \frac{\sqrt{A}}{1+\sqrt{A}}$$

$$\text{and} \quad V_{\min}\left\{\widehat{Y}_{2c}(st)\right\} = V_{opt}\left\{\widehat{Y}_{2c}(st) \middle| \lambda = \lambda_0\right\}$$

$$= \frac{1}{n}\left(\frac{1+\sqrt{A}}{2}\right) V_{y_2|z}$$

Remark 11.2.10

The relative efficiency of $\widehat{Y}_{2c}(st)$ compared to the Hansen–Hurwitz estimator based on a PPSWR sample of size n selected on the second occasion without using information on the first occasion is

$$E(st) = V\left(\widehat{Y}_{2n}\right) / V_{\min}\left\{\widehat{Y}_{2c}(st)\right\} = \frac{2}{1+\sqrt{A}}$$

The efficiency increases with the value of $|\delta_l|$ and it reaches the maximum 2 when $|\delta_l| = 1 \ \forall \ l = 1,\ldots, L$. Arnab (1991) showed that $V_{\min}\left\{\widehat{Y}_{2c}(st)\right\}$ is more efficient than the strategy proposed by Raj (1965a,b) and Chaudhuri and Arnab (1979a).

11.2.2 Sampling More Than Two Occasions

Consider a finite population $U = (u_1,\ldots, u_i,\ldots, u_N)$ of N identifiable units that is supposed to be surveyed up to $h(\geq 2)$ occasions. Let y_{ki} be the value of the ith unit of study variable y for the kth occasion, $i = 1,\ldots, N$; $k = 1,\ldots, h$. Let $Y_k = \sum_{i \in U} y_{ki}$ be the population total of y for the kth occasion.

11.2.2.1 Probability Proportional to Size With Replacement Sampling

Tripathi and Srivastava (1979) considered the following sampling scheme over k occasions. On the first occasion, a sample s_1 of size n is selected by PPSWR method using $p_i(>0)$ as normed size measure for the ith unit. On the second occasion a subsample s_{m_2} of size $m_2(\leq n)$ is selected from s_1 by SRSWOR method, treating all the n units in s_1 as distinct. Clearly, a unit can appear more than once in s_{m_2} if this unit had appeared more than once in s_1. The unmatched sample s_{u_2} of size $u_2(=n - m_2)$ is selected from the entire population by PPSWR method using normed size measures p_i's. The selected sample on the second occasion is denoted by $s_2 = s_{m_2} \cup s_{u_2}$. In general, at the $k(\geq 2)$th occasion, a matched sample s_{m_k} of size $m_k(\leq n)$ is selected by SRSWOR method from the sample s_{k-1}, selected at the $k - 1$th occasion. It is supplemented by an unmatched sample s_{u_k} of size $u_k(=n - m_k)$ selected by PPSWR method from the entire population U using p_i as normed size measures for the ith unit, $i = 1,\ldots, N$. The sample selected on the kth occasion is denoted by $s_k = s_{m_k} \cup s_{u_k}$ for $k = 2,\ldots, h$.

Theorem 11.2.10

The probability of selection of the jth ($j = 1,\ldots, N$) unit at the rth ($r = 1,\ldots, n$) draw for selecting the sample s_k is p_j where draws are labeled from 1 to m_k for the selection of the matched sample s_{m_k} and m_{k+1} through n for the selection of unmatched sample s_{u_k}.

Proof

Because the sample s_1 and the unmatched sample s_{u_k}'s are selected by PPSWR method directly from the entire population U, the probability of selection of the jth unit at any draw in s_1 and the also the probability of selection of the jth unit at the m_{k+1}th draw to nth draw in s_{u_k} are also equal to p_j.

For $r = 1, \ldots, m_k$ we note:

$Prob$ (selection of jth unit at rth draw in $s_{k+1} | j$th unit appeared q times in s_k)

$$= \sum_{x=0}^{r-1} Prob \text{ (jth unit selected x times in the first $(r-1)$ draws in } s_{k+1} | j\text{th unit appeared } q \text{ times in } s_k) \times \frac{q-x}{n-(r-1)}$$

$$= \sum_{x=0}^{r-1} \frac{\binom{q}{x}\binom{n-q}{r-1-x}}{\binom{n}{r-1}} \frac{q-x}{n-(r-1)}$$

$$= q/n \left(\text{Since } \frac{\sum_{x=0}^{r-1}\binom{q}{x}\binom{n-q}{r-1-x}}{\binom{n}{r-1}} = 1 \right.$$

$$\left. \text{and } \frac{\sum_{x=0}^{r-1} x\binom{q}{x}\binom{n-q}{r-1-x}}{\binom{n}{r-1}} = \frac{(r-1)q}{n} \right)$$

Hence the unconditional probability of selection of the jth unit at rth ($r \le m_2$) draw in s_2

$$= \sum_{q=0}^{n} Prob \text{ (selection of } j\text{th unit at } r\text{th draw in } s_2 | j\text{th unit appears } q \text{ times in } s_1) \times Prob \text{ (jth unit appears q times in } s_1)$$

$$= \sum_{q=0}^{n} \frac{q}{n} \left\{ \binom{n}{q} p_j^q (1-p_j)^{n-q} \right\}$$

$$= p_j$$

Hence we can show that the probability of selecting the jth unit at the rth draw in s_k is p_j by the method of induction.

11.2.2.1.1 Estimator for the Total on hth Occasion

The proposed estimator for $Y_h(\geq 2)$ is

$$\widehat{Y}_h = \phi_h \widehat{Y}_{hm} + (1-\phi_h)\widehat{Y}_{hu} \tag{11.2.37}$$

where $$\widehat{Y}_{hm} = \frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{hi}}{p_i} - \beta_h\left\{\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{h-1i}}{p_i} - \widehat{Y}_{h-1}(\min)\right\},$$

$\widehat{Y}_{hu} = \frac{1}{u_h}\sum_{i\in s_{u_h}}\frac{y_{hi}}{p_i}$, β_h is a constant chosen to minimize $V\left(\widehat{Y}_{hm}\right)$, and $\widehat{Y}_{h-1}(\min)$ is the estimator $\widehat{Y}_{h-1}$ with the minimum variance (obtained by choosing optimum values of ϕ_{h-1}, m_{h-1} and β_{h-1}).

Here, we make the following assumptions:

$$\begin{aligned} V_{k|z} &= \sum_{i\in U} p_i\left(\frac{y_{ki}}{p_i} - Y_k\right)^2 = V_0 \text{ and} \\ \delta_{k,k'} &= \sum_{i\in U} p_i\left(\frac{y_{k+1,i}}{p_i} - Y_{k+1}\right)\left(\frac{y_{k'i}}{p_i} - Y_{k'}\right)\Big/ V_o \\ &= \delta \quad \text{for } k\neq k';\, k,k' = 1,\ldots,h \end{aligned} \tag{11.2.38}$$

The optimum value of β_h for $h=2$ was obtained in the Theorem 11.2.4 as $\beta_2=\delta$. Hence for the sake of simplicity we assume $\beta_k=\delta$ for $k=1,\ldots,h$ and we write

$$\widehat{Y}^0_{hm} = \frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{hi}}{p_i} - \delta\left\{\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{h-1i}}{p_i} - \widehat{Y}_{h-1}(\min)\right\} \tag{11.2.39}$$

Theorem 11.2.11

(i) $E\left(\widehat{Y}^0_{hm}\right) = Y_h$

(ii) $V\left(\widehat{Y}^0_{hm}\right) = \left(1-\delta^2\right)\frac{V_0}{m_h} + \delta^2 V\left\{\widehat{Y}_{h-1}(\min)\right\}$

Proof

(i) From Theorem 11.2.10, we note that s_{m_2} and s_{u_2} are PPSWR samples from U and hence

$$E\left(\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{ki}}{p_i}\right)=E\left(\frac{1}{u_h}\sum_{i\in s_{u_h}}\frac{y_{ki}}{p_i}\right)=Y_k \quad \text{for } k=1,\ldots,h.$$

For $h=2$, the theorem follows from Theorem 11.2.4. Let us suppose that the theorem holds for $h=k-1$.

Then we have

$$\begin{aligned}E\left(\widehat{Y}^{0}_{km}\right)&=E\left(\frac{1}{m_k}\sum_{i\in s_{m_k}}\frac{y_{ki}}{p_i}\right)-\delta\left[E\left(\frac{1}{m_k}\sum_{i\in s_{m_k}}\frac{y_{k-1i}}{p_i}\right)-E\left\{\widehat{Y}_{k-1}(\min)\right\}\right]\\&=Y_k-\delta\left(Y_{k-1}-Y_{k-1}\right)=Y_k\end{aligned} \tag{11.2.40}$$

and

$$\begin{aligned}V\left(\widehat{Y}^{0}_{hm}\right)&=V\left(\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{hi}}{p_i}\right)+\delta^2 V\left\{\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{h-1i}}{p_i}-\widehat{Y}_{h-1}(\min)\right\}\\&\quad-2\delta Cov\left[\left(\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{hi}}{p_i}\right),\left(\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{h-1i}}{p_i}-\widehat{Y}_{h-1}(\min)\right)\right]\end{aligned} \tag{11.2.41}$$

Now, using the assumptions (Eq. 11.2.38) we get

$$V\left(\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{hi}}{p_i}\right)=\frac{V_o}{m_h}, \tag{11.2.42}$$

$$\begin{aligned}V\left\{\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{h-1i}}{p_i}-\widehat{Y}_{h-1}(\min)\right\}&=\frac{V_o}{m_h}+V\left\{\widehat{Y}_{h-1}(\min)\right\}\\&\quad-2Cov\left\{\frac{1}{m_h}\sum_{i\in s_{m_h}}\frac{y_{h-1i}}{p_i},\widehat{Y}_{h-1}(\min)\right\}\end{aligned} \tag{11.2.43}$$

and

$$Cov\left[\frac{1}{m_h}\sum\nolimits_{i\in s_{m_h}}\frac{y_{hi}}{p_i},\left\{\frac{1}{m_h}\sum\nolimits_{i\in s_{m_h}}\frac{y_{h-1i}}{p_i}-\widehat{Y}_{h-1}(\min)\right\}\right]$$

$$=\delta\frac{V_0}{m_h}-\delta Cov\left\{\frac{1}{m_h}\sum\nolimits_{i\in s_{m_h}}\frac{y_{h-1,i}}{p_i},\widehat{Y}_{h-1}(\min)\right\} \tag{11.2.44}$$

Putting Eqs. (11.2.42)–(11.2.44) in Eq. (11.2.41), we get

$$V\left(\widehat{Y}^0_{hm}\right)=\frac{V_o}{m_h}+\delta^2\left[\frac{V_o}{m_h}+V\left\{\widehat{Y}_{h-1}(\min)\right\}\right]-2\delta^2\frac{V_0}{m_h}$$

$$=\left(1-\delta^2\right)\frac{V_0}{m_h}+\delta^2 V\left\{\widehat{Y}_{h-1}(\min)\right\}$$

Hence the theorem is proved by the method of induction.

Theorem 11.2.12

The optimum proportion of matched sample $\lambda_h=\dfrac{m_h}{n}$ and minimum variance of $\widehat{Y}_h$ are, respectively, given by

$$\lambda_{h0}=\frac{\sqrt{1-\delta^2}}{1+\sqrt{1-\delta^2}}\frac{1}{g_{h-1}}\quad\text{and}$$

$$V_{\min}\left(\widehat{Y}_h\right)=V_{\min}(h)=\frac{\left(1+\sqrt{1-\delta^2}\right)g_{h-1}}{\left(1+\sqrt{1-\delta^2}\right)g_{h-1}+\left(1-\sqrt{1-\delta^2}\right)}\frac{V_0}{n}$$

where $g_h=\dfrac{V_{\min}(h)}{V_0/n}$ for $h\geq 2$ and $g_1=1$.

Proof

Minimizing the variance of $\widehat{Y}_h\left[=\phi_h\widehat{Y}^0_{hm}+(1-\phi_h)\widehat{Y}_{hu}\right]$ with respect to ϕ_h, the optimum value of ϕ_h is obtained as

$$\phi_{ho}=\frac{1}{V\left(\widehat{Y}^0_{hm}\right)}\left(\frac{1}{V\left(\widehat{Y}^0_{hm}\right)}+\frac{1}{V\left(\widehat{Y}_{hu}\right)}\right)^{-1}$$

Using Theorem 11.2.11, the optimum value of the variance of $\widehat{Y}_h$ with $\phi_h = \phi_{h0}$ is obtained as

$$V_{opt}\left(\widehat{Y}_h\right) = \left(\frac{1}{V\left(\widehat{Y}^0_{hm}\right)} + \frac{1}{V\left(\widehat{Y}_{hu}\right)}\right)^{-1}$$
$$= \left[\left\{\frac{\left(1-\delta^2\right)}{m_h} + \delta^2 \frac{V_{\min}(h-1)}{V_0}\right\}^{-1} + u_h\right]^{-1} V_0 \tag{11.2.45}$$

Now, minimizing $V_{opt}\left(\widehat{Y}^0_{hm}\right)$ with respect to m_h, we get the optimum value of λ_h as

$$\lambda_{ho} = \frac{\sqrt{1-\delta^2}}{1+\sqrt{1-\delta^2}} \frac{1}{g_{h-1}}$$

Substituting $m_h = n \ \lambda_{ho}$ in Eq. (11.2.45), we get the minimum variance of $\widehat{Y}_h$ as

$$V_{\min}(h) = \frac{\left(1+\sqrt{1-\delta^2}\right)g_{h-1}}{\left(1+\sqrt{1-\delta^2}\right)g_{h-1} + \left(1-\sqrt{1-\delta^2}\right)} \frac{V_0}{n}$$

Remark 11.2.11

The estimator $\widehat{Y}^0_{hm}$ cannot be used in practice because it involves an unknown parameter δ. So, in practice, we replace δ with its suitable estimate or some known value based on previous surveys.

When h is sufficiently large, we find $g = \underset{h\to\infty}{Lim}\, g_h = \dfrac{2\sqrt{1-\delta^2}}{1+\sqrt{1-\delta^2}}$, $\lambda = \underset{h\to\infty}{Lim}\, \lambda_{h0} = \frac{1}{2}$, and

$$V_{\min} = \underset{h\to\infty}{Lim}\, V\left\{\widehat{Y}_h(\min)\right\} = \frac{2\sqrt{1-\delta^2}}{1+\sqrt{1-\delta^2}} \frac{V_0}{n}.$$

Remark 11.2.12

The relative efficiency of $\widehat{Y}_h$ based on the data on previous $h-1$ occasions with respect to complete unmatched sample based on the data on hth occasion only is

$$E_h = \frac{V_0/n}{V\left\{\widehat{Y}_h(\min)\right\}} = 1 + \frac{1-\sqrt{1-\delta^2}}{\left(1+\sqrt{1-\delta^2}\right)g_{h-1}}$$

The efficiency $E_h \geq 1$ as $g_h \geq 0$.

11.2.2.2 Simple Random Sampling With Replacement

On the first occasion, a sample s_1 of size n is selected by SRSWR method. On the second occasion, a subsample s_{m_2} of size $m_2(\le n)$ is selected from s_1 by SRSWOR method treating all the units in s_1 as distinct. The unmatched sample s_{u_2} of size $u_2=(n-m_2)$ is selected from the entire population by SRSWR method. The selected sample on the second occasion is denoted by $s_2 = s_{m_2} \cup s_{u_2}$. In general, at the $k(\ge 2)$th occasion a matched sample s_{m_k} of size $m_k(\le n)$ is selected by SRSWOR method from the sample s_{k-1}, selected in the $k-1$th occasion and it is supplemented by an unmatched sample s_{u_k} of size $u_k(=n-m_k)$ selected by SRSWR method from the entire population U.

Because PPSWR sampling scheme reduces to SRSWR sampling when $p_i = 1/N$ for $i = 1,\ldots, N$, the assumption (Eq. 11.2.38) reduces to

$$\sigma^2_{ky} = \frac{1}{N}\sum_{i\in U}\left(y_{ki}-\overline{Y}_k\right)^2 = \sigma^2_y \quad \text{and}$$

$$\rho_{k,k'} = \frac{\sum_{i\in U}\left(y_{k,i}-\overline{Y}_k\right)\left(y_{k',i}-\overline{Y}_{k'}\right)}{\left(\sigma_{ky}\sigma_{k'y}\right)} \tag{11.2.46}$$

$$= \rho \quad \text{for} \quad k\neq k';\, k, k' = 1, \ldots, h$$

11.2.2.2.1 Estimator for the Total on the *h*th Occasion

Substituting $p_i = 1/N$ for PPSWR sampling scheme given in Section 11.2.2.1, we get $\delta = \rho$, $\widehat{Y}^0_{hm} = N\left[\overline{y}_h(s_{hm}) - \rho\left\{\overline{y}_{h-1}(s_{hm}) - \widehat{\overline{Y}}_{h-1}(\min)\right\}\right]$, $\widehat{Y}_{hu} = N\overline{y}(s_{uh})$, and

$$\widehat{Y}_h = \phi_h\widehat{Y}^0_{hm} + (1-\phi_h)\widehat{Y}_{hu} \tag{11.2.47}$$

where

$$\overline{y}(s_{uh}) = \sum_{i\in s_{hu}} y_{hi}/u_h,\ \overline{y}_h(s_{hm}) = \sum_{i\in s_{hm}} y_{hi}/m_h,\ \overline{y}_{h-1}(s_{hm}) = \sum_{i\in s_{hm}} y_{h-1,i}/m_h.$$

We have the following results, which follow immediately from Theorems 11.2.11 and 11.2.12 by substituting $p_i = 1/N$ for $i = 1,\ldots, N$.

Theorem 11.2.13

(i) $E\left(\widehat{Y}^0_{hm}\right) = Y_h$

(ii) $V\left(\widehat{Y}^0_{hm}\right) = N^2\left[\left(1-\rho^2\right)\frac{\sigma^2_y}{m_h} + \rho^2\frac{a_{h-1}}{n}\right]$

(iii) $V_{opt}\left(\widehat{Y}_h\right) = N^2\left[\left\{\dfrac{(1-\rho^2)}{m_h}+\rho^2\dfrac{a_{h-1}}{n}\right\}^{-1}+u_h\right]^{-1}\sigma_y^2$

(iv) $\lambda_{ho} = \dfrac{\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}}\dfrac{1}{a_{h-1}}$

(v) $V\left\{\widehat{Y}_h(\min)\right\} = V_{\min}(h) = N^2\dfrac{\left(1+\sqrt{1-\rho^2}\right)a_{h-1}}{\left(1+\sqrt{1-\rho^2}\right)a_{h-1}+\left(1-\sqrt{1-\rho^2}\right)}\dfrac{\sigma_y^2}{n}$

where $a_h = \dfrac{V_{\min}(h)}{N^2\sigma_y^2\Big/n}$ for $h \geq 2$ and $a_1 = 1$.

Remark 11.2.13

For large h, $a = \underset{h\to\propto}{Lim}\ a_h = \dfrac{2\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}}$, $\lambda = \underset{h\to\propto}{Lim}\ \lambda_h = 1\big/2$, and

$$V_{\min} = \underset{h\to\propto}{Lim}\ V\left\{\widehat{Y}_h(\min)\right\} = N^2\frac{2\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}}\frac{\sigma_y^2}{n}$$

11.3 ESTIMATION OF CHANGE OVER TWO OCCASIONS

For estimating the difference $D\left(=\overline{Y}_2-\overline{Y}_1\right)$ between the two means of the two occasions, we consider the following estimator:

$$\widehat{D} = \phi t_{2m} + \psi_1 t_{1m} + \psi_2\widehat{\widehat{Y}}_1 + \varphi\widehat{\widehat{Y}}_{2u} \tag{11.3.1}$$

where t_{2m}, t_{1m}, $\widehat{\widehat{Y}}_1$, $\widehat{\widehat{Y}}_{2u}$ are the same as in Section 11.2.1.2 and ϕ, ψ_1, ψ_2, and φ are suitably chosen to make $\widehat{D}$ unbiased for D. The unbiasedness condition $E\left(\widehat{D}\right) = D$ yields

$$\begin{aligned} D &= (\varphi+\phi)\overline{Y}_2 + (\psi_1+\psi_2)\overline{Y}_1 \\ \text{i.e.,}\quad \phi &= 1-\varphi \ \text{ and } \ \psi_1+\psi_2 = -1 \end{aligned} \tag{11.3.2}$$

Substituting Eq. (11.3.2) in Eq. (11.3.1), we get

$$\begin{aligned} \widehat{D} &= \phi\left(\widehat{\widehat{Y}}_{2m}-\widehat{\widehat{Y}}_1\right)+(1-\phi)\left(\widehat{\widehat{Y}}_{2u}-\widehat{\widehat{Y}}_1\right) \\ &= \phi\widehat{D}_1+(1-\phi)\widehat{D}_2 \end{aligned} \tag{11.3.3}$$

where $\widehat{D}_1 = \widehat{\widehat{Y}}_{2m}-\widehat{\widehat{Y}}_1$, $\widehat{D}_2 = \widehat{\widehat{Y}}_{2u}-\widehat{\widehat{Y}}_1$, $\widehat{\widehat{Y}}_{2m} = t_{2m}-\beta\left(t_{1m}-\widehat{\widehat{Y}}_1\right)$ with $\beta = -\psi_1/\phi$.

The optimum value of β that minimizes the variance of $\widehat{\widehat{Y}}_{2m}$ is $\beta = \beta_o = E\{Cov(t_{1m}, t_{2m}|s_1)\}/E\{V(t_{1m}|s_1)\}$.

The estimator $\widehat{\widehat{Y}}_{2m}$ with $\beta = \beta_o$ is

$$\widehat{\widehat{Y}}{}^{0}_{2m} = t_{2m}-\beta_0\left(t_{1m}-\widehat{\widehat{Y}}_1\right) \tag{11.3.4}$$

Substitution of $\hat{\overline{Y}}_{2m} = \hat{\overline{Y}}^{0}_{2m}$ in Eq. (11.3.4) yields

$$\hat{D} = \phi \hat{D}_{10} + (1-\phi)\hat{D}_2 \tag{11.3.5}$$

where $\hat{D}_{10} = \hat{\overline{Y}}^{0}_{2m} - \hat{\overline{Y}}_1$.

The optimum values of ϕ, $\hat{D}$, and the proportion of matched sample λ can be derived following the same procedure described in Section 11.2.1.2.

11.3.1 Simple Random Sampling Without Replacement

Let s_1, s_{2m}, and s_{2u} be selected by the SRSWOR method as in Section 11.2.1.3. In this situation

$$\hat{\overline{Y}}^{0}_{2m} = t_{2m} - \beta_0\left(t_{1m} - \hat{\overline{Y}}_1\right),\ \hat{\overline{Y}}_1 = \overline{y}_{1n},\ t_{2m} = \overline{y}_{2m},\ \hat{\overline{Y}}_{2u} = \overline{y}_{2u} \text{ and}$$

$$\beta_0 = S_{12y}\Big/ S^2_{1y}$$

The estimator $\hat{D}$ reduces to

$$\hat{D} = \phi \hat{D}_{10} + (1-\phi)\hat{D}_2 \tag{11.3.6}$$

with $\hat{D}_{10} = \left\{\overline{y}_{2m} - \beta_0\left(\overline{y}_{1m} - \overline{y}_{1n}\right)\right\} - \overline{y}_{1n}$ and $\hat{D}_2 = \overline{y}_{2u} - \overline{y}_{1n}$.

The variances of $\hat{D}_1, \hat{D}_2$ and covariance of $\hat{D}_1$ and $\hat{D}_2$ are given as follows:

$$\begin{aligned}
V\left(\hat{D}_{10}\right) &= V\left\{\overline{y}_{2m} - \beta_0\left(\overline{y}_{1m} - \overline{y}_{1n}\right)\right\} + V\left(\overline{y}_{1n}\right) - 2Cov\left[\left\{\overline{y}_{2m} - \beta_0\left(\overline{y}_{1m} - \overline{y}_{1n}\right)\right\}, \overline{y}_{1n}\right] \\
&= \left(\frac{1}{m} - \frac{1}{n}\right)(1-\rho^2)S^2_{2y} + \left(\frac{1}{n} - \frac{1}{N}\right)S^2_{2y} + \left(\frac{1}{n} - \frac{1}{N}\right)S^2_{1y} - 2Cov\left(\overline{y}_{2n}, \overline{y}_{1n}\right) \\
&= \left(\frac{1}{m} - \frac{1}{n}\right)(1-\rho^2)S^2_{2y} + \left(\frac{1}{n} - \frac{1}{N}\right)S^2_{2y} + \left(\frac{1}{n} - \frac{1}{N}\right)S^2_{1y} - 2\left(\frac{1}{n} - \frac{1}{N}\right)\rho S_{1y}S_{2y}
\end{aligned}$$

$$V\left(\hat{D}_2\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2_{1y} + \left(\frac{1}{u} - \frac{1}{N}\right)S^2_{2y} + \frac{2}{N}\rho S_{1y}S_{2y}$$

$$\begin{aligned}
Cov\left(\hat{D}_{10}, \hat{D}_2\right) &= Cov\left(\overline{y}_{2n} - \overline{y}_{1n}, \overline{y}_{2u}\right) - Cov\left(\overline{y}_{2n} - \overline{y}_{1n}, \overline{y}_{1n}\right) \\
&= -\frac{1}{N}\left(S^2_{2y} - \rho S_{1y}S_{2y}\right) - \left(\frac{1}{n} - \frac{1}{N}\right)\left(\rho S_{1y}S_{2y} - S^2_{1y}\right) \\
&= Q_{12}(\text{say})
\end{aligned}$$

The optimum value of ϕ that minimizes the variance of $\widehat{D}$ for given values of m and $u(=n-m)$ is

$$\phi_0 = (1/Q_1 + 1/Q_2)^{-1}/Q_1 \tag{11.3.7}$$

where $Q_1 = V\left(\widehat{D}_{10}\right) - Q_{12} = \left(\frac{1}{m} - \frac{1}{n}\right)(1-\rho^2)S_{2y}^2 + \frac{1}{n}\left(S_{2y}^2 - S_{12y}\right)$

and $Q_2 = V\left(\widehat{D}_2\right) - Q_{12} = \frac{1}{u}S_{2y}^2 + \frac{1}{n}S_{12y}$.

The variance of $\widehat{D}$ with $\phi = \phi_0$ is given by

$$V_{opt}\left(\widehat{D}\right) = (1/Q_1 + 1/Q_2)^{-1} + Q_{12}. \tag{11.3.8}$$

In particular, if $S_{1y}^2 = S_{2y}^2 = S_y^2$, we have

$$Q_1 = \left(\frac{1-\mu\rho^2}{\lambda} - \rho\right)\frac{S_y^2}{n},\ Q_2 = \left(\frac{1}{\mu} + \rho\right)\frac{S_y^2}{n} \text{ and}$$
$$Q_{12} = (1-2f)(1-\rho)S_y^2\Big/n \tag{11.3.9}$$

where $f = n/N$ and $\mu = 1-\lambda$.

After a little simplification we arrive at the following expression for $V_{opt}\left(\widehat{D}\right)$ as obtained by Särndal et al. (1992).

$$V_{opt}\left(\widehat{D}\right) = 2(1-\rho)\left(\frac{1}{1-\mu\rho} - f\right)\frac{S_y^2}{n} \tag{11.3.10}$$

From the expression (Eq. 11.3.10), we note that $V_{opt}\left(\widehat{D}\right)$ is an increasing function of μ when ρ is positive. So, our strategy will be to take a 100% matched sample on the second occasion to estimate the difference D if $\rho > 0$. In this case the minimum variance of $\widehat{D}$ is

$$V_{\min}\left(\widehat{D}\right) = 2(1-\rho)(1-f)\frac{S_y^2}{n} \tag{11.3.11}$$

11.4 ESTIMATION OF MEAN OF MEANS

For the estimation of the mean of means of two occasions $M = \left(\overline{Y}_1 + \overline{Y}_2\right)/2$, we may consider the following linear combinations as an unbiased estimator for M

$$\widehat{M} = \frac{1}{2}\left(\phi t_{2m} + \psi_1 t_{1m} + \psi_2 \widehat{\overline{Y}}_1 + \varphi \widehat{\overline{Y}}_{2u}\right) \tag{11.4.1}$$

The unbiasedness condition $E\left(\widehat{M}\right) = M$ implies $\varphi + \phi = \psi_1 + \psi_2 = 1$ and hence we can write

$$\widehat{M} = \frac{1}{2}\left(\phi\widehat{M}_1 + (1-\phi)\widehat{M}_2\right) \tag{11.4.2}$$

where $\widehat{M}_1 = \widehat{\widehat{Y}}_{2m} + \widehat{\widehat{Y}}_1$, $\widehat{\widehat{Y}}_{2m} = t_{2m} - \beta\left(t_{1m} - \widehat{\widehat{Y}}_1\right)$, $\beta = -\psi_1/\phi$, and $\widehat{M}_2 = \widehat{\widehat{Y}}_{2u} + \widehat{\widehat{Y}}_1$.

11.4.1 Simple Random Sampling Without Replacement

For SRSWOR sampling, $\widehat{M}_1 = \left\{\overline{y}_{2m} - \beta_0\left(\overline{y}_{1m} - \overline{y}_{1n}\right)\right\} + \overline{y}_{1n}$, $\beta_0 = S_{12y}\big/S_{1y}^2$, and $\widehat{M}_2 = \overline{y}_{2u} + \overline{y}_{1n}$. The variances of $\widehat{M}_1, \widehat{M}_2$ and covariance of $\widehat{M}_1$ and $\widehat{M}_2$ are given as follows:

$$V\left(\widehat{M}_1\right) = \left[\left(\frac{1}{m} - \frac{1}{n}\right)(1-\rho^2) + \left(\frac{1}{n} - \frac{1}{N}\right)\right]S_{2y}^2 + \left(\frac{1}{n} - \frac{1}{N}\right)S_{1y}^2 + 2\left(\frac{1}{n} - \frac{1}{N}\right)\rho S_{1y}S_{2y}$$

$$V\left(\widehat{M}_2\right) = \left[\left(\frac{1}{n} - \frac{1}{N}\right)S_{1y}^2 + \left(\frac{1}{u} - \frac{1}{N}\right)S_{2y}^2 - \frac{2}{N}\rho S_{1y}S_{2y}\right]$$

$$Cov\left(\widehat{M}_1, \widehat{M}_2\right) = -\frac{1}{N}\left(S_{2y}^2 + \rho S_{1y}S_{2y}\right) + \left(\frac{1}{n} - \frac{1}{N}\right)\left(\rho S_{1y}S_{2y} + S_{1y}^2\right) = \widetilde{Q}_{12}$$

In particular, if $S_{1y}^2 = S_{2y}^2 = S_y^2$, we have

$$\begin{aligned}
\widetilde{Q}_1 &= V\left(\widehat{M}_1\right) - Cov\left(\widehat{M}_1, \widehat{M}_2\right) \\
&= \left[\left(\frac{1}{m} - \frac{1}{n}\right)(1-\rho^2) + \frac{1}{n}(1+\rho)\right]S_y^2 \\
&= \frac{(1+\rho)}{n\lambda}(1-\rho\mu)S_y^2 \\
\widetilde{Q}_2 &= V\left(\widehat{M}_2\right) - Cov\left(\widehat{M}_1, \widehat{M}_2\right) = \left(\frac{1}{u} - \frac{1}{n}\right)\rho S_y^2 \\
&= \left(\frac{1-\mu\rho}{n\mu}\right)S_y^2
\end{aligned}$$

and

$$\widetilde{Q}_{12} = Cov\left(\widehat{M}_1, \widehat{M}_2\right) = \frac{(1-2f)}{n}(1+\rho)S_y^2$$

The variance of $\widehat{M}$ with the optimum value of ϕ is given by

$$\begin{aligned} V_{opt}\left(\widehat{M}\right) &= \frac{1}{4}\left[\left(1/\widetilde{Q}_1 + 1/\widetilde{Q}_2\right)^{-1} + \widetilde{Q}_{12}\right] \\ &= \frac{(1+\rho)}{2n}\left(\frac{1}{1+\rho\mu} - f\right)S_y^2 \end{aligned} \tag{11.4.3}$$

From the expression (Eq. 11.4.3), we note that $V_{opt}\left(\widehat{M}\right)$ is a deceasing function of μ when ρ is positive. Hence for positive ρ the optimum choice of the matched sample should be zero percent, i.e., an unmatched sample should be the same percent for estimation of M. In this case the minimum variance of $\widehat{M}$ is

$$V_{\min}\left(\widehat{M}\right) = \{1 - f(1+\rho)\}\frac{S_y^2}{2n} \tag{11.4.4}$$

Example 11.4.1

In the year 2008, a sample s_1 of 20 automobile dealers was selected from 50 dealers of a locality by the SRSWOR method. In 2009, a subsample s_{2m} of 8 dealers was selected from s_1 by the SRSWOR method and an unmatched sample of 12 dealers was selected from 30 dealers who were not selected in 2008 by SRSWOR method again. The sales of automobiles are given in the following table.

	Sales of automobiles			Sales of automobiles	
Dealers	**2008**	**2009**	**Dealers**	**2008**	**2009**
1	400	700	17	430	
2	350	300	18	225	
3	500	600	19	450	
4	625	725	20	700	
5	700	800	21		800
6	325	400	22		400
7	425	400	23		850

Dealers	Sales of automobiles 2008	2009	Dealers	Sales of automobiles 2008	2009
8	500	625	24		900
9	650		25		325
10	750		26		400
11	225		27		600
12	650		28		780
13	450		29		690
14	600		30		900
15	500		31		750
16	375		32		650

(i) Estimate the total sales of automobiles in the year 2009 along with its standard error. Estimate the optimum proportion of the matched sample for the second occasion if the total sample size is kept fixed at 20 for both of the occasions.
(ii) Estimate the change in the average sales of automobiles between these two periods with its standard error.
(iii) Estimate the average sales over two periods and its standard error.

Here, $N=50$, $n=20$, $m=8$, $u=12$; $\overline{y}_{1n}=491.5$, $\overline{y}_{1m}=478.125$, $\overline{y}_{2m}=568.750$, $\overline{y}_{2u}=670.417$; $s_{1m}^2=\sum_{i\in s_{2m}}\left(y_{1i}-\overline{y}_{1m}\right)^2/(m-1)$ $=17{,}220.982$, $s_{2m}^2=\sum_{i\in s_{2m}}\left(y_{2i}-\overline{y}_{2m}\right)^2/(m-1)=32{,}633.929$, $s_{2u}^2=\sum_{i\in s_{2u}}\left(y_{2i}-\overline{y}_{2u}\right)^2/(u-1)=40{,}374.811$, $s_{12m}=\dfrac{1}{m-1}\sum_{i\in s_{2m}}\left(y_{1i}-\overline{y}_{1m}\right)\left(y_{2i}-\overline{y}_{2m}\right)=19{,}308.035$, $\widehat{\rho}=s_{12m}/(s_{1m}s_{2m})=0.8144$, $\widehat{\beta}_0=\widehat{\rho}s_{2m}/s_{1m}=1.1212$, $s_{1y}^2=\sum_{i\in s_1}\left(y_{1i}-\overline{y}_1\right)^2/(n-1)=24{,}063.421$, $s_{2y}^2=\sum_{i\in s_2}\left(y_{2i}-\overline{y}_2\right)^2/(n-1)$ $=38{,}009.145$, $s_2=s_{2m}\cup s_{2u}$, and $\overline{y}_2=\sum_{i\in s_2}y_{2i}/n=629.75$.

(i) Estimated total sales in 2009 based on the matched sample is

$$\begin{aligned}\widehat{Y}_{2m}^{0}&=N\left\{\overline{y}_{2m}-\widehat{\beta}_0\left(\overline{y}_{1m}-\overline{y}_1\right)\right\}\\&=50\{568.750-1.1212(478.125-491.5)\}\\&=29{,}187.297\end{aligned}$$

Estimated total sales in 2009 based on unmatched sample

$$\widehat{Y}_{2u} = N\overline{y}_{2u} = 50 \times 670.416 = 33,520.8$$

Composite estimator for the total sales in 2009 is

$$\widehat{Y}_2 = \widehat{\phi}_0 \widehat{Y}^0_{2m} + \left(1 - \widehat{\phi}_0\right)\widehat{Y}_{2u}$$

where

$$\widehat{\phi}_0 = \frac{1/\hat{V}'_m}{1/\hat{V}'_m + 1/\hat{V}'_u} = \left\{\left(\frac{1}{m} - \frac{1}{n}\right)\left(1 - \widehat{\rho}^2\right) + \frac{1}{n}\right\}^{-1}$$

$$\left[\left\{\left(\frac{1}{m} - \frac{1}{n}\right)\left(1 - \widehat{\rho}^2\right) + \frac{1}{n}\right\}^{-1} + u\right]^{-1}$$

(using Eq. 11.2.22)
$= 0.526$
Hence estimated sales for the year 2009 is

$$\begin{aligned}\widehat{Y}_2 &= 0.526 \times 29,187.297 + (1 - 0.526) \times 33,520.833 \\ &= 31,243.592\end{aligned}$$

Estimated variance of $\widehat{Y}_2$ is

$$\widehat{V}\left(\widehat{Y}_2\right) = N^2\left[\left[\left\{\left(\frac{1}{m} - \frac{1}{n}\right)\left(1 - \widehat{\rho}^2\right) + \frac{1}{n}\right\}^{-1} + u\right]^{-1} - 1/N\right] s^2_{2y}$$

where $s^2_{2y} = \sum_{i\in s_2}\left(y_{2i} - \overline{y}_2\right)^2/(n-1) = 38,009.145$ is an unbiased estimate of S^2_{2y} based on the sample s_2.

Hence an estimated standard error of $\widehat{Y}_2$ is

$$Se\left\{\widehat{V}\left(\widehat{Y}_2\right)\right\} = \sqrt{\widehat{V}\left(\widehat{Y}_2\right)} = \sqrt{1,856,963.99} = 1362.704$$

Estimated optimum proportion of the matched sample is

$$\widehat{\lambda}_0 = \frac{\sqrt{1 - \widehat{\rho}^2}}{1 + \sqrt{1 - \widehat{\rho}^2}} = \frac{\sqrt{1 - 0.8144^2}}{1 + \sqrt{1 - 0.8144^2}} = 0.3672$$

(ii) Estimated average change of sales between the period 2008 and 2009 is

$$\widehat{D} = \widehat{\phi}_0 \widehat{D}_1 + \left(1 - \widehat{\phi}_0\right)\widehat{D}_2$$

Here,

$$\widehat{D}_1 = \left\{\overline{y}_{2m} - \widehat{\beta}_0\left(\overline{y}_{2m} - \overline{y}_{1n}\right)\right\} - \overline{y}_{1n} = \widehat{Y}^0_{2m}\Big/50 - \overline{y}_{1n}$$
$$= 583.746 - 491.5 = 92.246 \quad \text{and}$$
$$\widehat{D}_2 = \overline{y}_{2u} - \overline{y}_{1n} = 670.417 - 491.5 = 178.917;$$

$$\widehat{Q}_1 = \left(\frac{1}{m} - \frac{1}{n}\right)\left(1 - \widehat{\rho}^2\right)s^2_{2m} + \frac{1}{n}\left(s^2_{2y} - s_{12m}\right)$$
$$= (1/8 - 1/20)(1 - 0.8144^2) \times 32,633.929 + (38,009.145$$
$$-19,308.035)/20 = 1759.273,$$

$$\widehat{Q}_2 = \frac{1}{u}s^2_{2y} + \frac{1}{n}s_{12m} = 38,009.145/12 + 19,308.035/20 = 4132.831,$$

$$\widehat{Q}_{12} = \left(\frac{1}{n} - \frac{1}{N}\right)\left(s^2_{1m} - S_{12m}\right) - \frac{1}{N}\left(s^2_{2y} - s_{12m}\right) = (1/20 - 1/50)$$
$$(17,220.982 - 19,308.035) - (38,009.145 - 19,308.035)/50$$
$$= -436.633,$$

$$\widehat{\phi}_0 = \frac{1\big/\hat{Q}_1}{1\big/\hat{Q}_1 + 1\big/\hat{Q}_2} = 0.701$$

Hence $\widehat{D} = 0.701 \times 92.246 + 0.299 \times 178.917 = 118.161.$

Estimated standard error of $\widehat{D} = \sqrt{\widehat{V}\left(\widehat{D}\right)}$

$$= \sqrt{(1/Q_1 + 1/Q_2)^{-1} + Q_{12}} = 28.237.$$

(iii) The estimated average sale of automobiles for the year 2008 and 2009 is

$$\widehat{M} = \frac{1}{2}\left\{\widehat{\phi}_0\widehat{M}_1 + \left(1 - \widehat{\phi}_0\right)\widehat{M}_2\right\}$$

where $\widehat{M}_1 = \widehat{Y}^0_{2m}\Big/50 + \overline{y}_{1n} = 4583.746 + 91.5 = 1075.246;$

$$\widehat{M}_2 = \overline{y}_{2u} + \overline{y}_{1n} = 670.417 + 491.5 = 1161.917;$$

$$\widehat{V}\left(\widehat{M}_1\right) = \left[\left(\frac{1}{m} - \frac{1}{n}\right)\left(1 - \widehat{\rho}^2\right) + \left(\frac{1}{n} - \frac{1}{N}\right)\right]s^2_{2y} + \left(\frac{1}{n} - \frac{1}{N}\right)$$
$$\times s^2_{1y} + 2\left(\frac{1}{n} - \frac{1}{N}\right)s_{12m}$$
$$= \left[(1/8 - 1/20)\left(1 - 0.8144^2\right)\right.$$
$$\left. + (1/20 - 1/50)\right] \times 38,009.145 + (1/20 - 1/50)$$
$$\times 24,063.421 + 2(1/20 - 1/50) \times 19,308.035$$
$$= 3980.635;$$

$$\widehat{V}\left(\widehat{M}_2\right) = \left[\left(\frac{1}{n}-\frac{1}{N}\right)s_{1y}^2 + \left(\frac{1}{u}-\frac{1}{N}\right)s_{2y}^2 - \frac{2}{N}s_{12m}\right]$$
$$= (1/20 - 1/50) \times 24,063.421 + (1/12 - 1/50)$$
$$\times 38,009.145 - 2 \times 19,308.035/50 = 2356.827;$$
$$\widehat{\widehat{Q}}_{12} = \widehat{C}ov\left(\widehat{M}_1, \widehat{M}_2\right) = -\frac{1}{N}\left(s_{2y}^2 + s_{12m}\right) + \left(\frac{1}{n}-\frac{1}{N}\right)\left(s_{12m} + s_{1y}^2\right)$$
$$= -(38,009.145 + 19,308.035)/50 + (1/20 - 1/50)$$
$$(24,063.421 + 19,308.035)$$
$$= 154.800;$$
$$\widehat{\widehat{Q}}_1 = \widehat{V}\left(\widehat{M}_1\right) - \widehat{\widehat{Q}}_{12} = 3825.516, \quad \widehat{\widehat{Q}}_2 = \widehat{V}\left(\widehat{M}_2\right) - \widehat{\widehat{Q}}_{12} = 2202.027,$$
$$\widehat{\phi}_0 = 1/\widehat{\widehat{Q}}_1 / \left(1/\widehat{\widehat{Q}}_1 + 1/\widehat{\widehat{Q}}_2\right) = 0.365.$$
$$\widehat{M} = \frac{1}{2}\left\{\widehat{\phi}_0 \widehat{M}_1 + \left(1 - \widehat{\phi}_0\right)\widehat{M}_2\right\} = 565.141$$

Estimated standard error of $\widehat{M}$ is

$$SE\left(\widehat{M}\right) = \sqrt{\widehat{V}_{opt}\left(\widehat{M}\right)} = \sqrt{\frac{1}{4}\left(\frac{1}{1/\widehat{\widehat{Q}}_1 + 1/\widehat{\widehat{Q}}_2} + \widehat{\widehat{Q}}_{12}\right)} = 19.700$$

11.5 EXERCISES

11.5.1 On the first occasion a sample s_1 of size n is selected from a population U by RHC method of sampling with normed size measure p_i for the ith unit. On the second occasion a subsample s_{2m} of size $m(\le n)$ is chosen out of s_1 by the SRSWOR method and an independent sample s_{2u} is chosen from U by the RHC method again using the previous normed size measures p_i's. Let $\widehat{Y}_{2m} = \frac{n}{m}\sum_{i\in s_{2m}}\frac{y_{2i}-y_{1i}}{p_i}P_i + \sum_{i\in s_1}\frac{y_{1i}}{p_i}P_i$ and $\widehat{Y}_{2u} = \sum_{i\in s_{2u}}\frac{y_{2i}}{p_i}P_i^*$, where P_i and P_i^* denote the sum of the p_j's for the groups containing the ith unit in selecting samples s_{2m} and s_{2u} by RHC sampling schemes, respectively. Show that

(i) $\widehat{Y}_2 = \phi\widehat{Y}_{2m} + (1-\phi)\widehat{Y}_{2u}$ is an unbiased estimator for the population total Y_2 for any given value of ϕ,

(ii) the optimum proportion of unmatched sample $\mu(=1-\lambda)$ that minimizes the variance of $\widehat{Y}_2$ with the optimum value of ϕ is $\mu_0 = \left[1 + \sqrt{2(1-\delta)(1+\gamma f)}\,\right]^{-1}$, and

(iii) the minimum variance of $\widehat{Y}_2$ is $V_{\min}\left(\widehat{Y}_2\right) = \frac{N}{2(N-1)n}$ $\left[(1-f) + \sqrt{2(1-\delta)(1+\gamma f)}\,\right] V_0$

where $f = n/N$, $V_0 = \sum_{i \in U} p_i \left(\frac{y_{1i}}{p_i} - Y_1\right)^2 = \sum_{i \in U} p_i \left(\frac{y_{2i}}{p_i} - Y_2\right)^2$,

$\delta = \sum_{i \in U} p_i \left(\frac{y_{1i}}{p_i} - Y_1\right)\left(\frac{y_{2i}}{p_i} - Y_2\right) \Big/ V_0$,

$\gamma = \frac{(1-\rho)\overline{V}}{(1-\delta)V_0} - 1, \quad \overline{V} = N \sum_{i \in U} (y_{1i} - Y_1)^2 = N \sum_{i \in U} (y_{2i} - Y_2)^2$,

and $\rho = N \sum_{i \in U} (y_{1i} - Y_1)(y_{2i} - Y_2) \Big/ \overline{V}$ (Ghangurde and Rao, 1969; Chotai, 1974).

11.5.2 Let the initial sample s_1 and the unmatched sample s_{2u} be selected as in Exercise 11.5.1, but the matched sample s_{2m} is selected by RHC sampling replacing p_i by P_i defined in Exercise 11.5.1. Show that

(i) $\widehat{Y}_2 = \phi \widehat{Y}'_{2m} + (1-\phi)\widehat{Y}_{2u}$ is an unbiased estimator Y_2,

(ii) the optimum proportion of an unmatched sample with the optimum value of ϕ is $\mu_0 = \left[1 + \sqrt{2(1-\delta)}\,\right]^{-1}$, and

(iii) the minimum variance of $\widehat{Y}_2$ is $V_{\min}\left(\widehat{Y}_2\right) = \frac{N}{2(N-1)n}$ $\left[(1-f) + \sqrt{2(1-\delta)}\,\right] V_0$

where $f = n/N$, $\widehat{Y}'_{2m} = \sum_{i \in s_{2m}} \frac{y_{2i} - y_{1i}}{p_i} \Delta_i + \sum_{i \in s_1} \frac{y_{1i}}{p_i} P_i$, Δ_i be the sum of P_j-values from the group containing ith unit in selection of s_{2m} from s_1, and $\widehat{Y}_{2u}$, δ, and V_0 are as in Exercise 11.5.1 (Chotai, 1974).

11.5.3 On the first occasion, a sample s_1 of size n is selected from a population U by SRSWOR. On the second occasion, a matched sample s_{2m} of size m is selected from s_1 by the RHC sampling using normed size measure $p_i^* = y_{1i} \big/ \sum_{i \in s_1} y_{1i}$ for the unit $i \in s_1$ and an unmatched sample s_{2u} of size $u = n - m$ is selected from $U - s_1$. Let $\widehat{\overline{Y}}_{2m} = \frac{1}{n} \sum_{i \in s_{2m}} \frac{y_{2i}}{p_i} P_i, \quad p_i = y_{1i}/Y_1, \quad \widehat{\overline{Y}}_{2u} = \frac{1}{u} \sum_{i \in s_{2u}} y_{2i}$, and $\widehat{\overline{Y}}_2 = \phi \widehat{\overline{Y}}_{2m} + (1-\phi)\widehat{\overline{Y}}_{2u}$, where P_i is the sum of p_j-values of the group containing the ith unit in selecting sample s_{2m} by the RHC procedure. Show that

(i) $E\left(\widehat{\overline{Y}}_{2m}\right) = \overline{Y}_2$,

(ii) $V\left(\widehat{\overline{Y}}_{2m}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{2y}^2 + \left(\frac{1}{m} - \frac{1}{n}\right) \frac{1}{N(N-1)}$

$\sum\limits_{i \in U} \left(y_{2i} - R^* y_{1i}\right)^2,$

(iii) $Cov\left(\widehat{\overline{Y}}_{2m}, \widehat{\overline{Y}}_{2u}\right) = -S_{2y}^2/N,$

(iv) the optimum proportion of the unmatched sample which minimizes the variance of $\widehat{\overline{Y}}_2 = \phi \widehat{\overline{Y}}_{2m} + (1-\phi)\widehat{\overline{Y}}_{2u}$ with the optimum value of ϕ is $\mu_0 = 1/\left(1 + \sqrt{\delta^*}\right)$, and

(v) the minimum variance of $\widehat{\overline{Y}}_2$ is $V_{\min}\left(\widehat{\overline{Y}}_2\right) = \left(\frac{1+\sqrt{\delta^*}}{2n} - \frac{1}{N}\right) S_{2y}^2$

where $R^* = Y_2/Y_1$ and $\delta^* = \dfrac{\sum\limits_{i \in U} \left(y_{2i} - R^* y_{1i}\right)^2 / p_i}{N \sum\limits_{i \in U} \left(y_{2i} - \overline{Y}_2\right)^2}$ (Avadhan and Sukhatme, 1970).

11.5.4 On the first occasion, a sample s_1 of size n is selected from a population U by the SRSWOR method. On the second occasion, a subsample s_{2m} of size m is selected from s_1 by the SRSWOR method and an unmatched sample s_{2u} of size $u(=n-m)$ is selected from the entire population U by SRSWOR. Let $s_{2u} = s_{21} \cup s_{22}$ where $s_{21} = s_{2m} \cap s_{2u}$ denote the units of s_{2u} that come from the match sample and s_{22} be the remaining units of s_{2u}. Let $t = \phi \overline{y}_{2u} + (1-\phi)\overline{y}'_{2m}$ and $t^* = \phi \dfrac{m_2 \overline{y}_{2m} + l_2 \overline{y}_{22}}{m_2 + l_2} + (1-\phi)\,\overline{y}'_{2m}$ where $\overline{y}_{2u} = \sum\limits_{i \in s_{2u}} y_{2i}/u$, $\overline{y}_{2m} = \sum\limits_{i \in s_{2m}} y_{2i}/m$, $\overline{y}'_{2m} = \overline{y}_{2m} - b\left(\overline{y}_{1m} - \overline{y}_1\right)$, $\overline{y}_1 = \sum\limits_{i \in s_1} y_{1i}/n$, $\overline{y}_{22} = \sum\limits_{i \in s_{22}} y_{2i}/l_2$, m_2 is the size of s_{21} and $l_2 = u - m_2$, and b is chosen to minimize the variance of $\overline{y}'_{2m}$. Show that

(i) $E(t) = E\left(t^*\right)$ and (ii) $V(t) - V\left(t^*\right) = \dfrac{m-1}{N-1}\left(\dfrac{1}{n-m} - \dfrac{1}{N}\right) \geq 0$ (Pathak and Rao, 1967).

11.5.5 Consider the Raj (1965a,b) sampling scheme where on the first occasion, a sample s_1 of size n is selected from a population U by the PPSWR method using the normed size measure p_i attached to the ith unit. On the second occasion, a matched sample s_{2m}

of size $m(<n)$ is selected from s_1 by the SRSWOR method treating all the units of s_1 are distinct and an unmatched sample s_{2u} of size $u = n - m$ units from the entire population by the PPSWR method using p_i as normed size measure for the ith unit. Let $t_0 = \phi t_m + (1-\phi)t_u$ and $t_1 = \phi t_m + (1-\phi)t_u^*$, where $t_m = \frac{1}{m}\sum_{i\in s_{2m}} \frac{y_{2i}-y_{1i}}{p_i} + \frac{1}{n}\sum_{i\in s_1} \frac{y_{1i}}{p_i}$, $t_u = \frac{1}{u}\sum_{i\in s_{2u}} \frac{y_{2i}}{p_i}$, $t_u^* = \frac{1}{u}\left(\sum_{i\in s_{22}} \frac{y_{2i}}{p_i} + m_2 \sum_{i\in s_{2m}} y_{2i} \Big/ \sum_{i\in p_{2m}} p_i\right)$, s_{22} is the sample units of s_{2u} that do not belong to s_{2m}, $\sum_{i\in s}$ = sum over the units in s including repetition, and $\sum_{i\in s}$ = sum over the distinct units in s. Show that (i) $E(t_0) = E(t_1) = \overline{Y}_2 = \sum_{i\in U} y_{2i}/N$ and (ii) $V(t_1) \le E(t_0)$ (Singh, 1972).

11.5.6 Consider the sampling scheme over two occasions in Exercise 11.5.2. If we modify the estimator $\widehat{Y}'_{2m}$ by $\widehat{Y}^0_{2m} = \sum_{i\in s_{2m}} \frac{y_{2i}-ky_{1i}}{p_i}\Delta_i + k\sum_{i\in s_1} \frac{y_{1i}}{p_i}P_i$ where k is a suitably chosen constant. Find the optimum value of k and the minimum variance of $\widehat{Y}_2 = \phi\widehat{Y}'_{2m} + (1-\phi)\widehat{Y}_{2u}$, where $\widehat{Y}_{2u}$ as in Exercise 11.5.2 (Chaudhuri and Arnab, 1979a).

11.5.7 On the first occasion, a sample s_1 of size n is selected from a population U of size n by the SRSWOR method. On the second occasion, a matched-sample s_{2m} of size $m(<n)$ is selected from s_1 by the SRSWOR method and an unmatched sample s_{2u} of size u is selected from $U - s_1$ by SRSWOR method. Suppose we want to estimate the population mean $\overline{Y}_1$. Find the optimum values of α_1, α_2, α_3, and α_4 for which the estimator

$$\widehat{\overline{Y}}_1 = \frac{\alpha_1}{n}\sum_{i\in s_1} y_{1i} + \frac{\alpha_2}{m}\sum_{i\in s_{2m}} y_{1i} + \frac{\alpha_3}{m}\sum_{i\in s_{2m}} y_{2i} + \frac{\alpha_4}{u}\sum_{i\in s_{2u}} y_{2i}$$

will be unbiased for $\overline{Y}_1$ and has a minimum variance for a given value of m. Find also the optimum value of m that minimizes the variance of $\widehat{\overline{Y}}_1$ and also obtain the expression for the minimum variance of $\widehat{\overline{Y}}_1$.

11.5.8 To estimate the average sales of a certain product in a region, an independent sample of size n was selected from each establishment by SRSWR method and information of the sales for the current and previous month were collected each month. Let $\bar{y}_h$ and $\bar{x}_{h-1}$ be the sample means of sales on hth and $h-1$ occasion based on the sample selected on the hth occasion. Let the proposed estimator of the mean sales on the hth occasion based on the data collected in the current and previous occasion be of the form $\widehat{\bar{Y}}_h = \bar{y}_h - \alpha_h \bar{x}_{h-1} + \alpha_h \widehat{\bar{Y}}_{h-1}$, where α_h is a suitably chosen constant and $\alpha_1 = 0$. Assuming the variance of sales of each of the occasions and correlation of sales between any two consecutive occasions remain the same as σ^2 and ρ, respectively, show that the minimum variance of $\widehat{\bar{Y}}_h$ is $V_{\min}\left(\widehat{\bar{Y}}_h\right) = (1-\rho\alpha_h)\sigma^2/n$, where α_h is obtained from $\alpha_h = \rho/(2-\rho\alpha_{h-1})$ (Ecler, 1955; Woodroff, 1959).

11.5.9 In 2005, a sample of 30 clinics was selected from a total of 75 clinics in a certain locality by SRSWOR method and the number of HIV/AIDS affected persons treated in the clinics was collected. In 2008, a subsample of 10 clinics was selected from the selected 30 clinics in 2005; in addition, a fresh sample of 20 clinics was selected from the 45 clinics that were not selected in 2005. The data are given below.

	No. of HIV/AIDS persons treated			No. of HIV/AIDS persons treated	
Clinics	**2005**	**2008**	**Clinics**	**2005**	**2008**
1	700	500	26	970	
2	850	800	27	1200	
3	1200	900	28	1870	
4	650	700	29	980	
5	950	800	30	750	
6	750	600	31		550
7	1500	1200	32		750
8	2500	1750	33		1010
9	600	400	34		850
10	300	420	35		600
11	750		36		325
12	700		37		750
13	875		38		895
14	900		39		750
15	750		40		650

Clinics	No. of HIV/AIDS persons treated 2005	2008	Clinics	No. of HIV/AIDS persons treated 2005	2008
16	870		41		900
17	1200		42		875
18	280		43		750
19	250		44		480
20	870		45		750
21	600		46		560
22	825		47		420
23	750		48		375
24	890		49		175
25	475		50		600

Based on the data collected in 2005 and 2008 estimate the following:

(i) Total number of HIV/AIDS affected persons treated in the locality in the year 2008 and its standard error.

(ii) Total number of HIV/AIDS affected persons treated in the locality in the year 2005 and its standard error.

(iii) Change in number of HIV/AIDS affected persons treated between the year 2005 and 2008 along with its standard error.

(iv) Total number of HIV/AIDS affected persons treated in the period between 2005 and 2008 altogether and its standard error.