A CONTRIBUTION TO DOUBLE SAMPLING

By Balkrishna V. Sukhatme

Institute of Agricultural Research Statistics, New Delhi

and

R. S. Koshal

F.A.O., Cairo

[Received, August 1959]

1. Introduction

In sample surveys, the unknown mean $\bar{Y}$ of a finite population can often be estimated more efficiently by using an auxiliary variable X which is correlated with Y and whose mean $\bar{X}$ is known. One such estimate is the ratio estimate. Many a time it happens, however, that the population mean $\bar{X}$ is not known and in this case the ratio estimate cannot be used to estimate the unknown population mean $\bar{Y}$. The usual procedure in such a situation is to use what is known as the technique of double sampling. This paper is concerned with the extension of double sampling technique to multi-stage designs and illustrate the results with reference to data on paddy collected in the course of crop-cutting experiments conducted in Mansura District of Egypt.

2. Notation and Double Sampling Procedure

Let

N = total number of primary units in the population,

M_i = the number of secondary units in the i-th primary unit $(i = 1, 2, \ldots, N)$,

P_{ij} = the number of tertiary units in the ij-th secondary unit $(j = 1, 2, \ldots, M_i, \quad i = 1, 2, \ldots, N)$

Y_{ijk} = the value of the k-th tertiary unit of the j-th secondary unit of the i-th primary unit,

$$\bar{Y}_{ij.} = \frac{1}{P_{ij}} \sum_{k=1}^{P_{ij}} Y_{ijk},$$

$$\bar{Y}_{i..} = \frac{1}{\sum_{1}^{M_i} P_{ij}} \sum_{1}^{M_i} \sum_{1}^{P_{ij}} Y_{ijk} = \frac{1}{M_i \bar{P}_{i.}} \sum_{1}^{M_i} P_{ij} \bar{Y}_{ij.} = \frac{1}{M_i} \sum_{1}^{M_i} V_{ij} \bar{Y}_{ij.},$$

$$\bar{Y}_{...} = \frac{1}{\sum_{1}^{N} \sum_{1}^{M_i} P_{ij}} \sum_{1}^{N} \sum_{1}^{M_i} \sum_{1}^{P_{ij}} Y_{ijk},$$

$$= \frac{1}{\sum_{1}^{N} M_i \bar{P}_{i.}} \sum_{1}^{N} M_i \bar{P}_{i.} \bar{Y}_{i..} = \frac{1}{\sum_{1}^{N} Q_i} \sum_{1}^{N} Q_i \bar{Y}_{i..} = \frac{1}{N} \sum_{1}^{N} u_i \bar{Y}_{i..},$$

$$\bar{P}_{i.} = \frac{1}{M_i} \sum_{1}^{M_i} P_{ij}, \quad V_{ij} = \frac{P_{ij}}{\bar{P}_{i.}}, \quad Q_i = M_i \bar{P}_{i.},$$

where

$$\bar{Q} = \frac{1}{N} \sum_{1}^{N} Q_i \quad \text{and} \quad u_i = \frac{Q_i}{\bar{Q}}.$$

Further, let X_{ijk} be the value of the auxiliary variable for the k-th tertiary unit of the j-th secondary unit of the i-th primary unit. Then the various qualities $\bar{X}_{ij.}$, $\bar{X}_{i..}$ and $\bar{X}_{...}$ can be defined similarly as above.

When the population mean $\bar{X}_{...}$ is known, the usual estimator for estimating the population mean $\bar{Y}_{...}$ is

$$\bar{Y}_R = \frac{\hat{\bar{Y}}_{...}}{\hat{\bar{X}}_{...}} \cdot \bar{X}_{...}$$

where $\hat{\bar{Y}}_{...}$ and $\hat{\bar{X}}_{...}$ are the sample estimates of the populations means $\bar{Y}_{...}$ and $\bar{X}_{...}$ respectively.

We shall now consider the case when $\bar{X}_{...}$ is not known. In this case, we propose to use the procedure of double sampling. To estimate $\bar{X}_{...}$, we take a random sample of n' primary units out of N. In the i-th selected primary unit, we take a random sample of m_i' secondary units out of M_i. In the ij-th selected secondary unit, we select a random sample of p_{ij}' tertiary units of P_{ij}. Then an unbiased estimator of the population mean $\bar{X}_{...}$ is given by

$$\bar{X}'_{n'} = \frac{1}{n'} \sum_{1}^{n'} u_i \hat{\bar{X}}_{i..} \tag{2.1}$$

where

$$\hat{\bar{X}}_{i..} = \frac{1}{m_i'} \sum_{1}^{m_i'} V_{ij} \bar{X}_{ij}(p_{ij}'),$$

and its variance is given by

$$V(\bar{X}'_{n'}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_{bx}'^2 + \frac{1}{n'N} \sum_{1}^{N} u_i^2 \left(\frac{1}{m_i'} - \frac{1}{M_i}\right) S_{ix}'^2$$

$$+ \frac{1}{n'N} \sum_{1}^{N} \frac{u_i^2}{m_i' M_i} \sum_{1}^{M_i} V_{ij}^2 \left(\frac{1}{p_{ij}'} - \frac{1}{P_{ij}}\right) S^2_{ijx} \tag{2.2}$$

where

$$\left.\begin{aligned} S_{bx}'^2 &= \frac{1}{N-1} \sum_{1}^{N} (u_i \bar{X}_{i..} - \bar{X}_{...})^2 \\ S_{ix}'^2 &= \frac{1}{M_i - 1} \sum_{1}^{M_i} (V_{ij} \bar{X}_{ij.} - \bar{X}_{i..})^2 \end{aligned}\right\} \tag{2.3}$$

and

$$S'^2_{ijx} = \frac{1}{P_{ij} - 1} \sum_{1}^{P_{ij}} (X_{ijk} - \bar{X}_{ij.})^2$$

Out of n' primary units, we select a random sub-sample of n units. Out of the m_i' secondary units selected in the i-th selected primary unit, we select a random sub-sample of m_i secondary units. Out of the p_{ij}' tertiary units selected in the ij-th selected secondary unit, we select a random sub-sample of p_{ij} units. Then the total number of ultimate units selected in the sub-sample is

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} p_{ij}.$$

Both the characters X and Y are observed on this sub-sample. Then the estimate proposed for the population mean of the character Y, using information on the auxiliary variable X, is

$$\bar{Y}'_R = \frac{\bar{Y}_n}{\bar{X}_n} \cdot \bar{X}'_{n'} \tag{2.4}$$

with

$$\bar{Y}_n = \frac{1}{n}\sum_{1}^{n} u_i \hat{\bar{Y}}_{i..}$$

$$\hat{\bar{Y}}_{i..} = \frac{1}{m_i}\sum_{1}^{m_i} V_{ij}\, \bar{Y}_{ij(p_{ij})}$$

$$\bar{X}_n = \frac{1}{n}\sum_{1}^{n} u_i \hat{\bar{X}}_{i..}$$

$$\hat{\bar{X}}_{i..} = \frac{1}{m_i}\sum_{1}^{m_i} V_{ij}\, \bar{X}_{ij(p_{ij})}.$$

3. Expectation and Variance of $\bar{Y}'_R$

To compute the expectation of $\bar{Y}_R'$, we proceed as follows. Since $\bar{Y}_n$, $\bar{X}_n$ and $\bar{X}'_{n'}$ are unbiased estimates of the corresponding population means, we assume that $\bar{Y}_n = \bar{Y}_{...} + E_1$, $\bar{X}_n = \bar{X}_{...} + E_2$ and $\bar{X}'_{n'} = \bar{X}_{...} + E_2'$ with $EE_1 = EE_2 = EE_2' = 0$.

Then

$$E\bar{Y}'_R = E\frac{\bar{Y}_n}{\bar{X}_n}\bar{X}'_{n'}$$

$$= E\frac{(\bar{Y}_{...} + E_1)(\bar{X}_{...} + E_2')}{(\bar{X}_{...} + E_2)}.$$

Expanding, taking expectation and neglecting higher order terms it can be shown that

$$E(\bar{Y}_R') = \bar{Y}_{...}(1 + B_1) \tag{3.1}$$

where

$$B_1 = \frac{\mathrm{Cov}(\bar{Y}'_{n'}, \bar{X}'_{n'}) - \mathrm{Cov}(\bar{Y}_n, \bar{X}_n)}{\bar{Y}_{...}\bar{X}_{...}} - \frac{V(\bar{X}'_{n'}) - V(\bar{X}_n)}{\bar{X}^2_{...}}. \tag{3.2}$$

Hence the estimate is biased, the relative bias of the estimate being given by (3.2).

Taking P_{ij}, M_i and N to be very large so that finite correction factors can be ignored and further letting $m_i' = m'$, $m_i = m$, $p_{ij}' = p'$, $p_{ij} = p$, $S_{ixy} = \bar{S}_{wxy}$ and $S_{ijxy} = \bar{\bar{S}}_{wxy}$, it turns out that

$$B_1 = \left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{S_{bx}^2}{\bar{\bar{X}}_{...}^2} - \frac{S_{bxy}}{\bar{\bar{X}}_{...}\bar{\bar{Y}}_{...}}\right)$$

$$+ \left(\frac{1}{nm} - \frac{1}{n'm'}\right)\left(\frac{\bar{S}_{wx}^2}{\bar{\bar{X}}_{...}^2} - \frac{\bar{S}_{wxy}}{\bar{\bar{X}}_{...}\bar{\bar{Y}}_{...}}\right)$$

$$+ \left(\frac{1}{nmp} - \frac{1}{n'm'p'}\right)\left(\frac{\bar{\bar{S}}_{wx}^2}{\bar{\bar{X}}_{...}^2} - \frac{\bar{\bar{S}}_{wxy}}{\bar{\bar{X}}_{...}\bar{\bar{Y}}_{...}}\right) \qquad (3.3)$$

It follows that the bias will be zero and hence the estimate unbiased if

$$\frac{\bar{\bar{Y}}_{...}}{\bar{\bar{X}}_{...}} = \frac{S_{bxy}}{S_{bx}^2} = \frac{\bar{S}_{wxy}}{\bar{S}_{wx}^2} = \frac{\bar{\bar{S}}_{wxy}}{\bar{\bar{S}}_{wx}^2}. \qquad (3.4)$$

Using the results of large sample theory, we have

$$V(\bar{Y}_R') = V(\bar{Y}_n) + R^2\,[V(\bar{X}_n) - V(X'_{n'})]$$
$$- 2R\,[\mathrm{Cov}(\bar{Y}_n, \bar{X}_n) - \mathrm{Cov}(\bar{Y}'_{n'}, \bar{X}'_{n'})] \qquad (3.5)$$

where

$$R = \frac{\bar{\bar{Y}}_{...}}{\bar{\bar{X}}_{...}}.$$

Substituting and simplifying, we obtain

$$V(\bar{Y}_R') = \left(\frac{1}{n} - \frac{1}{n'}\right)(S_{by}'^2 + R^2 S_{by}'^2 - 2RS'_{bxy})$$

$$+ \frac{1}{nN}\sum_1^N u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)(S_{iy}'^2 + R^2 S_{ix}'^2 - 2RS'_{ixy})$$

$$- \frac{1}{n'N}\sum_1^N u_i^2\left(\frac{1}{m_i'} - \frac{1}{M_i}\right)(S_{iy}'^2 + R^2 S_{ix}'^2 - 2RS'_{ixy})$$

$$+ \frac{1}{nN}\sum_1^N \frac{u_i^2}{m_i M_i}\sum_1^{M_i} V_{ij}^2\left(\frac{1}{p_{ij}} - \frac{1}{P_{ij}}\right)$$
$$\times (S^2_{ijy} + R^2 S^2_{ijx}) - 2RS_{ijxy})$$

$$- \frac{1}{n'N}\sum_1^N \frac{u_i^2}{m_i' M_i}\sum_1^{M_i} V_{ij}^2\left(\frac{1}{p_{ij}'} - \frac{1}{P_{ij}}\right)$$
$$\times (S^2_{ijy} + R^2 S^2_{ijx} - 2RS_{ijxy})$$

$$+\left(\frac{1}{n'}-\frac{1}{N}\right)S_{by}{}'^2+\frac{1}{n'N}\sum_1^N u_i^2\left(\frac{1}{m_i'}-\frac{1}{M_i}\right)S_{iy}{}'^2$$

$$+\frac{1}{n'N}\sum_1^N\frac{u_i^2}{m_i'M_i}\sum_1^{M_i}V_{ij}{}^2\left(\frac{1}{p_{ij}'}-\frac{1}{P_{ij}}\right)S^2{}_{ijy}. \qquad (3.6)$$

If $p_{ij}=p_{ij}'=P_{ij}$, that is, there is no sub-sampling at the third stage:

$$V(\bar{Y}_R')=\left(\frac{1}{n}-\frac{1}{n'}\right)(S_{by}{}'^2+\mathrm{R}^2S_{bx}{}'^2-2RS'_{bxy})$$

$$+\frac{1}{nN}\sum_1^N u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)(S_{iy}{}'^2+R^2S_{ix}{}'^2-2RS'_{ixy})$$

$$-\frac{1}{n'N}\sum_1^N u_i^2\left(\frac{1}{m_i'}-\frac{1}{M_i}\right)(S_{iy}{}'^2+R^2S_{ix}{}'^2-2RS'_{ixy})$$

$$+\left(\frac{1}{n'}-\frac{1}{N}\right)\mathrm{S}_{by}{}'^2+\frac{1}{n'N}\sum_1^N u_i^2\left(\frac{1}{m_i'}-\frac{1}{M_i}\right)S_{iy}{}'^2. \qquad (3.7)$$

If in addition, $m_i'=m_i=M_i$, that is, there is no sub-sampling at the second stage also:

$$V(\bar{Y}_R')=\left(\frac{1}{n'}-\frac{1}{N}\right)\mathrm{S}_{by}{}'^2+\left(\frac{1}{n}-\frac{1}{n'}\right)(S_{by}{}'^2+R^2S_{bx}{}'^2-2RS'_{bxy}). \qquad (3.8)$$

Let

$$\bar{S}_{wxy}=\frac{1}{N}\sum_1^N S_{ixy},\quad \bar{S}_{wx}{}^2=\frac{1}{N}\sum_1^N S_{ix}{}^2,\quad \bar{S}_{wy}{}^2=\frac{1}{N}\sum_1^N S_{iy}{}^2,$$

$$\bar{\bar{S}}_{wxy}=\frac{1}{NM}\sum_1^N\sum_1^M S_{ijxy},\quad \bar{\bar{S}}_{wx}{}^2=\frac{1}{NM}\sum_1^N\sum_1^M S^2{}_{ijx},$$

$$\bar{\bar{S}}_{wy}{}^2=\frac{1}{NM}\sum_1^N\sum_1^M S^2{}_{ijy},\quad S_{bq}{}^2=S_{by}{}^2+R^2S_{bx}{}^2-2RS_{bxy},$$

$$\bar{S}_{wq}{}^2=\bar{S}_{wy}{}^2+R^2\bar{S}_{wx}{}^2-2R\bar{S}_{wxy},$$

$$\bar{\bar{S}}_{wq}{}^2=\bar{\bar{S}}_{wy}{}^2+R^2\bar{\bar{S}}_{wx}{}^2-2R\bar{\bar{S}}_{wxy}.$$

If further it is assumed that $u_i = 1$, $V_{ij} = 1$, $m_i = m$, $m_i' = m'$, $p_{ij} = p$, $p_{ij}' = p'$ and finite correction factors can be ignored, formulæ 3.6, 3.7 and 3.8 reduce respectively to

$$V(\bar{Y}_R') = \left(\frac{1}{n} - \frac{1}{n'}\right) S_{bq}^2 + \left(\frac{1}{nm} - \frac{1}{n'm'}\right) \bar{S}_{wq}^2 + \left(\frac{1}{nmp} - \frac{1}{n'm'p'}\right) \bar{\bar{S}}_{wq}^2 + \frac{S_{by}^2}{n'} + \frac{\bar{S}_{wy}^2}{n'm'} + \frac{\bar{\bar{S}}_{wy}^2}{n'm'p'}. \tag{3.9}$$

$$V(\bar{Y}_R') = \left(\frac{1}{n} - \frac{1}{n'}\right) S_{bq}^2 + \left(\frac{1}{mn} - \frac{1}{m'n'}\right) \bar{S}_{wq}^2 + \frac{S_{by}^2}{n'} + \frac{\bar{S}_{wy}^2}{n'm'}. \tag{3.10}$$

$$V(\bar{Y}_R') = \left(\frac{1}{n} - \frac{1}{n'}\right) S_{bq}^2 + \frac{S_{by}^2}{n'}. \tag{3.11}$$

Formulæ (3.10) and (3.11) agree with those given by Hora[2] and Cochran.[1]

4. Estimation of the Variance

Let

$$s_{ijxy} = \frac{1}{p_{ij} - 1} \sum_{1}^{p_{ij}} (X_{ijk} - \bar{X}_{ij(p_{ij})})(Y_{ijk} - \bar{Y}_{ij(p_{ij})})$$

$$s'_{ixy} = \frac{1}{m_i - 1} \sum_{1}^{m_i} (V_{ij}\bar{X}_{ij(p_{ij})} - \bar{X}_{i(m_i)(p_{ij})})(V_{ij}\bar{Y}_{ij(p_{ij})} - \bar{Y}_{i(m_i)(p_{ij})})$$

$$s'_{bxy} = \frac{1}{n - 1} \sum_{1}^{n} (u_i\bar{X}_{i(m_i)(p_{ij})} - \bar{X}_n)(u_i\bar{Y}_{i(m_i)(p_{ij})} - \bar{Y}_n) \tag{4.1}$$

with s^2_{ijy}, s^2_{ijx}, $s'_{ix}{}^2$, $s_{iy}{}'^2$, $s_{bx}{}'^2$, $s_{by}{}'^2$ defined similarly.

Then it can be easily shown that

$$Es_{ijxy} = S_{ijxy}$$

$$Es'_{ixy} = S'_{ixy} + \frac{1}{M_i} \sum_{1}^{M_i} V_{ij}^2 \left(\frac{1}{p_{ij}} - \frac{1}{P_{ij}}\right) S_{ijxy}$$

$$Es'_{bxy} = S'_{bxy} + \frac{1}{N}\sum_{1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S'_{ixy}$$

$$+ \frac{1}{N}\sum_{1}^{N} \frac{u_i^2}{m_i M_i} \sum_{1}^{M_i} V_{ij}^2 \left(\frac{1}{p_{ij}} - \frac{1}{P_{ij}}\right) S_{ijxy}.$$

It follows that unbiased estimates of S_{ijxy}, S'_{ixy} and S'_{bxy} are given by:

$$\text{Est } S_{ijxy} = s_{ijxy} \tag{4.2}$$

$$\text{Est } S'_{ixy} = s'_{ixy} - \frac{1}{m_i}\sum_{1}^{m_i} V_{ij}^2 \left(\frac{1}{p_{ij}} - \frac{1}{P_{ij}}\right) s_{ijxy} \tag{4.3}$$

$$\text{Est } S'_{bxy} = s'_{bxy} - \frac{1}{n}\sum_{1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s'_{ixy}$$

$$- \frac{1}{n}\sum_{1}^{n} \frac{u_i^2}{m_i M_i} \sum_{1}^{m_i} V_{ij}^2 \left(\frac{1}{p_{ij}} - \frac{1}{P_{ij}}\right) s_{ijxy}. \tag{4.4}$$

Estimates for S^2_{ijx}, S^2_{ijy}, $S_{ix}{}'^2$, $s_{iy}{}'^2$, $S_{bx}{}'^2$ and $S_{by}{}'^2$ can be obtained n a similar manner. Also, ignoring bias

$$\text{Est } R = \frac{\bar{y}_n}{\bar{x}_n} = R_s \tag{4.5}$$

Using these estimates, it follows on simplification that

$$\text{Est } V(\bar{y}_R{}') = \left(\frac{1}{n} - \frac{1}{n'}\right)(s_{by}{}'^2 + R_s^2 s_{bx}{}'^2 - 2R_s s'_{bxy})$$

$$+ \frac{1}{nn'}\sum_{1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{m_i'}\right)(s_{iy}{}'^2 + R_s^2 s_{ix}{}'^2 - 2R_s s'_{ixy})$$

$$+ \left(\frac{1}{n'} - \frac{1}{N}\right) s_{by}{}'^2 + \frac{1}{nN}\sum_{1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{iy}{}'^2$$

$$- \frac{1}{nn'}\sum_{1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{m_i'}\right) s_{iy}{}'^2$$

$$+\frac{1}{nN}\sum_{1}^{n}\frac{{u_i}^2}{m_i M_i}\sum_{1}^{m_i}{V_{ij}}^2\left(\frac{1}{p_{ij}}-\frac{1}{P_{ij}}\right){s^2}_{ijy}$$

$$-\frac{1}{nn'}\sum_{1}^{n}\frac{{u_i}^2}{m_i m_i'}\sum_{1}^{m_i}{V_{ij}}^2\left(\frac{1}{p_{ij}}-\frac{1}{p_{ij}'}\right){s^2}_{ijy}. \qquad (4.6)$$

If all the $M_i's$ are equal to M say and P_{ij} equal to P and $m_i = m$, $m_i' = m'$, $p_i = p$ and $p_i' = p'$, then it can be shown that

$$E\bar{\bar{s}}_{wxy} = E\frac{1}{nm}\sum_{1}^{n}\sum_{1}^{m}s_{ijxy} = \bar{\bar{S}}_{wxy}$$

$$E\bar{s}^2_{wxy} = E\frac{1}{n}\sum_{1}^{n}s_{ixy} = \bar{S}_{wxy} + \frac{1}{p}\bar{\bar{S}}_{wxy}$$

$$Es_{bxy} = S_{bxy} + \frac{1}{m}\bar{S}_{wxy} + \frac{1}{mp}\bar{\bar{S}}_{wxy}$$

It follows that

$$\text{Est } \bar{\bar{S}}_{wxy} = \bar{\bar{s}}_{wxy} \qquad (4.7)$$

$$\text{Est } \bar{S}_{wxy} = \bar{s}_{wxy} - \frac{\bar{\bar{s}}_{wxy}}{p} \qquad (4.8)$$

$$\text{Est } S_{bxy} = s_{bxy} - \frac{\bar{s}_{wxy}}{m}. \qquad (4.9)$$

Writing

$${s_{bq}}^2 = {s_{by}}^2 + {R_s}^2{s_{bx}}^2 - 2R_s s_{bxy}$$
$${\bar{s}_{wq}}^2 = {\bar{s}_{wy}}^2 + {R_s}^2{\bar{s}_{wx}}^2 - 2R_s\bar{s}_{wxy}$$
$${\bar{\bar{s}}_{wq}}^2 = {\bar{\bar{s}}_{wy}}^2 + {R_s}^2{\bar{\bar{s}}_{wx}}^2 - 2R_s\bar{\bar{s}}_{wxy}$$

and ignoring finite correction factors, we find that the estimate of the variance given in (3.9) is given by:

$$\text{Est } V(\bar{Y}_R') = \left(\frac{1}{n}-\frac{1}{n'}\right){s_{bq}}^2 + \frac{1}{n'}\left(\frac{1}{m}-\frac{1}{m'}\right){\bar{s}_{wq}}^2$$
$$+\frac{1}{m'n'}\left(\frac{1}{p}-\frac{1}{p'}\right){\bar{\bar{s}}_{wq}}^2 + \frac{{s_{by}}^2}{n'}$$
$$-\frac{1}{n'}\left(\frac{1}{m}-\frac{1}{m'}\right){\bar{s}_{wy}}^2 - \frac{1}{m'n'}\left(\frac{1}{p}-\frac{1}{p'}\right){\bar{\bar{s}}_{wy}}^2. \qquad (4.10)$$

5. Efficiency

When no information is available concerning the auxiliary variable, the appropriate estimate for the population mean is $\bar{Y}_n$ and its variance is given by

$$V(\bar{Y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{by}'^2 + \frac{1}{nN} \sum_1^N u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{iy}'^2$$

$$+ \frac{1}{nN} \sum_1^N \frac{u_i^2}{m_i M_i} \sum_1^{M_i} V_{ij}^2 \left(\frac{1}{p_{ij}} - \frac{1}{P_{ij}}\right) S^2_{ijy} \qquad (5.1)$$

$$= \frac{S_{by}^2}{n} + \frac{\bar{S}_{wy}^2}{mn} + \frac{\bar{\bar{S}}_{wy}^2}{mnp}$$

if $m_i = m$, $p_{ij} = p$, $M_i = M$ and $P_{ij} = P$ and finite correction factors can be ignored.

Hence, double sampling will be more efficient than simple random sampling if

$$V(\bar{Y}_R') < V(\bar{Y}_n)$$

i.e., if

$$\left(\frac{1}{n} - \frac{1}{n'}\right)(R^2 S_{bx}^2 - 2RS_{bxy})$$

$$+ \left(\frac{1}{nm} - \frac{1}{n'm'}\right)(R^2 \bar{S}_{wx}^2 - 2R\bar{S}_{wxy})$$

$$+ \left(\frac{1}{nmp} - \frac{1}{n'm'p'}\right)(R^2 \bar{\bar{S}}_{wx}^2 - 2R\bar{\bar{S}}_{wxy}) < 0. \qquad (5.2)$$

If

$$\rho_b = \frac{S_{bxy}}{S_{bx} S_{by}}, \quad \bar{\rho}_w = \frac{\bar{S}_{wxy}}{\bar{S}_{wx} \bar{S}_{wy}}, \quad \bar{\bar{\rho}}_w = \frac{\bar{\bar{S}}_{wxy}}{\bar{\bar{S}}_{wx} \bar{\bar{S}}_{wy}},$$

$$C_{bx} = \frac{S_{bx}}{\bar{X}_{...}}, \quad C_{by} = \frac{S_{by}}{\bar{Y}_{...}}, \quad \bar{C}_{wx} = \frac{\bar{S}_{wx}}{\bar{X}_{..}}, \quad \bar{C}_{wy} = \frac{\bar{S}_{wy}}{\bar{Y}_{...}},$$

$$\bar{\bar{C}}_{wx} = \frac{\bar{\bar{S}}_{wx}}{\bar{X}_{...}} \quad \text{and} \quad \bar{\bar{C}}_{wy} = \frac{\bar{\bar{S}}_{wy}}{\bar{Y}_{...}}.$$

Then $V(\bar{Y}_R')$ is less than $V(\bar{Y}_n)$ if

$$\rho_b > \frac{1}{2}\frac{C_{bx}}{C_{by}}; \quad \bar{\rho}_w > \frac{1}{2}\cdot\frac{\bar{C}_{wx}}{\bar{C}_{wy}}, \quad \bar{\bar{\rho}}_w > \frac{1}{2}\cdot\frac{\bar{\bar{C}}_{wx}}{\bar{\bar{C}}_{wy}}.$$

In particular, if $C_{bx} = C_{by}$, $\bar{C}_{wx} = \bar{C}_{wy}$ and $\bar{\bar{C}}_{wx} = \bar{\bar{C}}_{wy}$, the estimate $\bar{Y}_R'$ based on double sampling will be more efficient than $\bar{Y}_n$ if $\rho_b > \frac{1}{2}$, $\bar{\rho}_w > \frac{1}{2}$ and $\bar{\bar{\rho}}_w > \frac{1}{2}$.

6. Numerical Illustration

The results obtained above will now be illustrated with reference to the data collected from the crop-cutting experiments on paddy carried out in the Mansura District of Egypt during the year 1955. The design adopted for the survey was that of multi-stage stratified sampling with villages constituting the primary units of sampling and fields as the secondary units of sampling. The data relate to weight in kilograms of grain and straw together and grain alone. If X denotes the weight of grain and straw together and Y the weight of grain alone, then from the analysis of the data, we obtain:

Analysis of variance

Source	D.F.	M.S. (y^2)	M.S. (yx)	M.S. (x^2)
Between villages	11	65·91	141·68	418·50
Within villages ..	12	8·21	28·12	153·26

Hence Est $\bar{S}_{wx}^2 = 153{\cdot}26$, Est $\bar{S}_{wy}^2 = 8{\cdot}21$, Est $\bar{S}_{wxy} = 28{\cdot}12$, Est $S_{bx}^2 = 132.62$, Est $S_{by}^2 = 28.85$, Est $S_{bxy} = 56.78$ and $R_s = {\cdot}2979$.

We will now consider different schemes of sampling

I	$n' = 48$	$m' = 1$
	$n = 12$	$m = 1$
II	$n' = 24$	$m' = 2$
	$n = 12$	$m = 1$
III	$n' = 16$	$m' = 3$
	$n = 12$	$m = 1$
IV	$n' = 12$	$m' = 4$
	$n = 12$	$m = 1$

Then using the above data, we have $V(\bar{Y}_R')_{\text{I}} = 1{\cdot}51$, $V(\bar{Y}_R')_{\text{II}} = 1{\cdot}97$, $V(\bar{Y}_R')_{\text{III}} = 2{\cdot}43$ and $V(Y_R')_{\text{IV}} = 2{\cdot}89$. Also $V(\bar{Y}_{mn})$ for $n = 12$, $m = 1$, without the use of auxiliary variable is $3{\cdot}09$. Therefore the gain in efficiency under the first scheme is 104% while that under the second and third schemes is 57% and 27% respectively. The gain in efficiency under the fourth scheme is only 7%.

These comparisons are of considerable value. What is of interest however is to know whether the reduction in variance would be worth the extra expenditure on the additional sample required to observe the auxiliary variable. Clearly, we should choose that procedure which for a fixed cost, say C_0, minimizes the variance of the estimate. Let

c_1 = cost of visiting a primary unit,

c_2 = cost per experiment of harvesting the crop and observing x, the yield of (grain + straw),

c_3 = cost per experiment of winnowing and thrashing the produce.

If we consider the second or third scheme of double sampling, the total cost C_0 can be expressed as

$$C_0 = c_1 n' + c_2 n' m' + c_3 n m \tag{6.1}$$

We will now determine the optimum values of n', n, m' and m such that for a fixed cost C_0, $V(\bar{Y}_R')$ is minimum. We therefore consider the function

$$V(Y_R') + \lambda (c_1 n' + c_2 n' m' + c_3 n m) \tag{6.2}$$

where $V(\bar{Y}_R')$ is given by (3.10). Then differentiating w.r.t. n', n, m' and m, we find after some simplification that the optimum values of n', n, m' and m are given by

$$n' = \frac{BC_0}{\sqrt{c_1}\,[B\sqrt{c_1} + W\sqrt{c_2} + \bar{S}_{wg}\sqrt{c_3}]},$$

$$m' = \frac{W\sqrt{c_1}}{B\sqrt{c_2}}, \quad m = 1 \tag{6.3}$$

and

$$n = \frac{\bar{S}_{wg} C_0}{\sqrt{c_3}\,[B\sqrt{c_1} + W\sqrt{c_2} + \bar{S}_{wg}\sqrt{c_3}]}$$

with

$$B^2 = S_{by}^2 - S_{bq}^2 \quad \text{and} \quad W^2 = \bar{S}_{wy}^2 - \bar{S}_{wq}^2$$

whence substituting, the optimum variance is given by

$$V(\bar{Y}_R') = \frac{1}{C_0}\left[B\sqrt{c_1} + W\sqrt{c_2} + \bar{S}_{wq}\sqrt{c_3} + \frac{S_{bq}^2}{\bar{S}_{wq}}\sqrt{c_3}\right] \times (B\sqrt{c_1} + W\sqrt{c_2} + \bar{S}_{wq}\sqrt{c_3}). \tag{6.4}$$

In the case of simple two-stage sampling design without the use of auxiliary variable, the appropriate cost function to be considered is

$$C_0 = c_1 n + (c_2 + c_3)\, mn \tag{6.5}$$

We will therefore determine the optimum values of m and n so that $V(\bar{Y}_{mn})$ is minimum for fixed cost C_0 with

$$V(\bar{Y}_{nm}) = \frac{S_{by}^2}{n} + \frac{\bar{S}_{wy}^2}{mn}. \tag{6.6}$$

Considering the function

$$V(\bar{Y}_{nm}) + \lambda(c_1 n + c_2 nm + c_3 mn) \tag{6.7}$$

and differentiating w.r.t. n and m, we find that the optimum values of n and m are given by

$$n = \frac{S_{by} C_0}{\sqrt{c_1}\,[\sqrt{c_1}\, S_{by} + \sqrt{c_2 + c_3}\, \bar{S}_{wy}]},$$

$$m = \frac{\bar{S}_{wy}\sqrt{c_1}}{\bar{S}_{by}\sqrt{c_2 + c_3}} \tag{6.8}$$

and that the optimum variance is given by

$$V(\bar{Y}_{nm}) = \frac{1}{C_0}[\sqrt{c_1}\, S_{by} + \sqrt{c_2 + c_3}\, \bar{S}_{wy}]^2. \tag{6.9}$$

Substituting for S_{by}^2, $\bar{S}_{wy}^2$, B^2, W^2, $\bar{S}_{wq}^2$ and S_{bq}^2 the estimates derived from the analysis of variance, a comparison was made of the two systems of sampling for $c_3/c_2 = 1$ and different values of c_1/c_2. These results are given below.

c_1/c_2 ..	9	16	25	36
% Gain in efficiency	6	11	14	16

If we consider the fourth scheme of double sampling, the total cost C_0 can be expressed as

$$C_0 = c_1 n + c_2 nm' + c_3 nm \tag{6.10}$$

Proceeding as before, we find that the optimum values of n, m' and m are given by

$$n = \frac{S_{by}\, C_o}{\sqrt{c_1}\,(\sqrt{c_1}\, S_{by} + \sqrt{c_2}\, W + \sqrt{c_3}\, \bar{S}_{wq}]},$$

$$m' = \frac{W\sqrt{c_1}}{S_{by}\sqrt{c_2}}$$

and

$$m = \frac{\bar{S}_{wq}\sqrt{c_1}}{S_{by}\sqrt{c_3}} \tag{6.11}$$

and that the optimum variance is given by

$$V(\bar{Y}_R') = \frac{[\sqrt{c_1}\, S_{by} + \sqrt{c_2}\, W + \sqrt{c_3}\, S_{wq}]^2}{C_0}. \tag{6.12}$$

Then the relative efficiency of double sampling over simple two stage sampling is given by

$$R.E. = \left[\frac{\sqrt{c_1}\, S_{by} + \sqrt{c_2 + c_3}\, \bar{S}_{wy}}{\sqrt{c_1}\, S_{by} + \sqrt{c_2}\, W + \sqrt{c_3}\, \bar{S}_{wq}}\right]^2. \tag{6.13}$$

Under this set-up it can be shown that double sampling will be more efficient than simple two-stage sampling, if

$$\frac{c_3}{c_2} > \frac{\bar{S}_{wq}^2}{W^2}. \tag{6.14}$$

If we consider the first scheme of double sampling, the total cost C_0 can be expressed as

$$C_0 = c_1 n' + c_2 n' m + c_3 nm \tag{6.15}$$

Under this set-up, it is not possible to obtain explicitly the optimum values of n', m and n and hence the optimum variance.

7. Stratified Sampling

The extension of the theory to stratified sampling is straightforward. For simplicity we will consider the simple case when each

primary unit contains the same number of secondary units and each secondary unit contains the same number of tertiary units. We will also use the same notation that was employed for the unstratified case except that an additional subscript t will be introduced to denote the stratum. Two procedures are usually followed:

(*i*) *Separate ratio estimate.*—A separate estimate is made for each stratum. Then the estimate of the population mean over-all strata is given by

$$\bar{Y}_R{}' = \sum_1^K \lambda_t \bar{Y}_{Rt}{}' \tag{7.1}$$

where $\lambda_t = (N_t/N)$ and K is the number of strata and $\sum_1^K \lambda_t = 1$. Then ignoring finite correction factors, it can be shown that

$$E\bar{Y}_R{}' = \sum_1^K \lambda_t \bar{Y}_{t...}\Big[1 + \left(\frac{1}{n_t} - \frac{1}{n_t'}\right)(C^2_{tbx} - \rho_{tb}\, C_{tbx} C_{tby}) + \left(\frac{1}{n_t m_t} - \frac{1}{n_t' m_t'}\right)(\bar{C}^2_{twx} - \bar{\rho}_{tw}\, \bar{C}_{twx}\, \bar{C}_{twy}) + \left(\frac{1}{n_t m_t p_t} - \frac{1}{n_t' m_t' p_t'}\right)(\bar{\bar{C}}^2_{twx} - \bar{\bar{\rho}}_{tw}\, \bar{\bar{C}}_{twx}\, \bar{\bar{C}}_{twy}) \tag{7.2}$$

To obtain an idea of how the bias diminishes with the sample size, assume that

$$n_t' = \alpha_1 n_t,\ m_t' = \alpha_2 m_t,\ p_t' = \alpha_3 p_t$$

where α_1, α_2 and α_3 are constants greater than unity and

$$n_t = \frac{n}{K},\ m_t = \frac{m}{K},\ p_t = \frac{p}{K}$$

where

$$\Sigma n_t = n,\ \Sigma m_t = m \text{ and } \Sigma p_t = p.$$

Further assume that the quantities C_{tbx}, C_{tby}, ρ_{tb}, $\bar{C}_{twx}$, $\bar{C}_{twy}$, $\bar{\rho}_{tw}$, $\bar{\bar{C}}_{twx}$, $\bar{\bar{C}}_{twy}$ and $\bar{\bar{\rho}}_{tw}$ are each of the same order from stratum to stratum, say C_{bx}, C_{by}, ρ_b, $\bar{C}_{wx}$, $\bar{C}_{wy}$, $\bar{\rho}_w$, $\bar{\bar{C}}_{wx}$, $\bar{\bar{C}}_{wy}$ and $\bar{\bar{\rho}}_w$ respectively. Then it will be seen that the relative bias in the estimate tends to zero provided the sample size within each stratum is sufficiently large.

Also, to a first approximation

$$V(\bar{Y}_R') = \sum_1^K \lambda_t^2 [V(\bar{Y}_{nt}) + R_t^2 [V(\bar{X}_{nt}) - V(\bar{X}_{n't})] - 2R_t [\mathrm{Cov}(\bar{Y}_{nt}, \bar{X}_{nt}) - \mathrm{Cov}(\bar{Y}_{n't}, \bar{X}_{nt})] \quad (7.3)$$

This formula is based on the assumption that the sample size in each stratum is sufficiently large. This may not always be true. We therefore consider the second procedure.

(ii) Combined estimate.—The estimate in this case is

$$\bar{Y}_{Rc}' = \frac{\sum_1^K \lambda_t \bar{Y}_{nt}}{\sum_1^K \lambda_t \bar{X}_{nt}} \cdot \sum_1^K \lambda_t \bar{X}_{n't}. \quad (7.4)$$

Under the assumptions made earlier and assuming in addition $n_t \sim N_t$, it can be shown that the relative bias in the estimate equals

$$\frac{1}{n}\left(1 - \frac{1}{\alpha_1}\right)(C_{bx}^2 - \rho_b C_{bx} C_{by}) + \frac{K}{nm}\left(1 - \frac{1}{\alpha_1\alpha_2}\right)(\bar{C}_{wx}^2 - \bar{\rho}_w \bar{C}_{wx} \bar{C}_{wy}) + \frac{K^2}{nmp}\left(1 - \frac{1}{\alpha_1\alpha_2\alpha_3}\right)(\bar{\bar{C}}_{wx}^2 - \bar{\bar{\rho}}_w \bar{\bar{C}}_{wx} \bar{\bar{C}}_{wy}).$$

It follows that even when the size of the sample within each stratum is small, a combined estimate can give a satisfactory estimate of the population mean provided the total sample n is sufficiently large.

To a first approximation, it can be shown that

$$V(\bar{Y}_{Rc}') = \sum_1^K \lambda_t^2 \{V(\bar{Y}_{nt}) + R^2 [V(\bar{X}_{nt}) - V(\bar{X}_{n't})] - 2R [\mathrm{Cov}(\bar{Y}_{nt}, \bar{X}_{nt}) - \mathrm{Cov}(\bar{Y}_{n't}, \bar{X}_{n't})]\}. \quad (7.6)$$

Then it follows that

$$V(\bar{Y}_{Rc}') - V(\bar{Y}_R') = \sum_1^K \lambda_t^2 \{[V(\bar{X}_{nt}) - V(\bar{X}_{nt'})](R - R_t)^2 + 2(R - R_t)[R_t \{V(\bar{X}_{nt}) - V(\bar{X}_{n't})\} - \{\mathrm{Cov}(\bar{X}_{nt}, \bar{Y}_{nt}) - \mathrm{Cov}(\bar{X}_{n't}, \bar{Y}_{n't})\}]\}.$$

It will be seen that the difference depends upon the magnitude of the variation between the strata ratios and the value of

$$R_t [V(\bar{X}_{nt}) - V(\bar{X}_{nt'})] - [\text{Cov}(\bar{X}_{nt}, \bar{Y}_{nt}) - \text{Cov}(\bar{X}_{n't}, \bar{Y}_{n't})].$$

The latter expression however represents the relative bias of the estimated population mean within each stratum which will generally be small. It follows therefore that the combined estimate will have a lower precision than that based on separate strata. On the other hand, the bias in the former estimate will be smaller than in the latter. Unless therefore the population ratios in the different strata vary considerably, the use of a combined estimate may provide an estimate which has a negligible bias and whose precision is almost as high as that of the estimate based on separate strata.

The results given in this paper have also been extended to the case when the selection probabilities are unequal. It is proposed to give these results in another paper.

8. Summary

For estimating the population total of a character y under study by means of ratio method of estimation, it is usually assumed that the population total of the ancillary character x is known. When this is not so, the usual practice known as double sampling is to take a large sample for estimating the population total of x and a sub-sample of this sample is used to observe both the characters. Cochran gives the appropriate formulæ in this case for single stage designs. This paper extends these results to multistage designs. The results are illustrated with reference to a sample survey on paddy.

9. Acknowledgment

The authors are grateful to Dr. V. G. Panse for several suggestions and constant encouragement in the preparation of this paper.

References

1. Cochran, W. G. .. *Sampling Techniques*, John Wiley and Sons, New York, 1953.

2. Hora, R. B. .. "On some extensions in double sampling," unpublished thesis for Diploma, I.C.A.R., 1956.