

Ratio Estimation with Subsampling the Nonrespondents

PODURI S.R.S. RAO¹

ABSTRACT

The procedure of subsampling the nonrespondents suggested by Hansen and Hurwitz (1946) is considered. Post-stratification prior to the subsampling is examined. For the mean of a characteristic of interest, ratio estimators suitable for different practical situations are proposed and their merits are examined. Suitable ratio estimators are also suggested for the situations in which the Hard-Core are present.

KEY WORDS: Auxiliary information; Post-stratification; Biases; Mean square errors; Linear model; Hard-Core.

1. INTRODUCTION

Consider a finite population of size N and a random sample of size n drawn without replacement. In surveys on human populations, frequently n_1 units respond on the items under examination, but the remaining $(n - n_1)$ units do not provide any response. The initial survey may be conducted through the mail or telephone calls, perhaps computer-aided.

In Sections 2, 3 and 4, we consider Hansen and Hurwitz's (1946) procedure of subsampling a portion of the $(n - n_1)$ nonrespondents. In this procedure the population is supposed to be consisting of the response stratum of size N_1 and the nonresponse stratum of size $N_2 = (N - N_1)$.

In Section 2, we discuss two procedures for post-stratifying the sampled units, prior to the subsampling of the nonrespondents.

Two ratio estimators for the mean of an item are considered in Section 3. Biases and Mean Square errors of these estimators are compared in Sections 3 and 4. In Section 4, two more ratio estimators, which may be suitable for some practical situations, are proposed and their relative merits are examined.

The Hard-Core problem is considered in Section 5. Six different estimators for this situation are proposed. Optimum conditions suitable for each one of the estimators are briefly described.

2. HANSEN AND HURWITZ'S ESTIMATOR AND POST-STRATIFICATION

Consider a characteristic of interest y_i , $i = (1, 2, \dots, N)$. Let $\bar{Y} = (\sum_1^N y_i)/N$ and $S^2 = \sum_1^N (y_i - \bar{Y})^2 / (N - 1)$ denote the mean and variance of the population. Let $\bar{Y}_1 = (\sum_1^{N_1} y_i)/N_1$ and $S_1^2 = \sum_1^{N_1} (y_i - \bar{Y}_1)^2 / (N_1 - 1)$ denote the mean and variance of the response group. Similarly, let $\bar{Y}_2 = (\sum_1^{N_2} y_i) / N_2$ and $S_2^2 = \sum_1^{N_2} (y_i - \bar{Y}_2)^2 / (N_2 - 1)$ denote the mean and variance of the nonresponse group. The population

¹ P.S.R.S. Rao, Department of Statistics, University of Rochester, Rochester, NY 14627, U.S.A.

mean can be written as $\bar{Y} = W_1 \bar{Y}_1 + W_2 \bar{Y}_2$, where $W_1 = (N_1/N)$ and $W_2 = (N_2/N)$. The sample mean $\bar{y}_1 = (\sum_1^n y_i)/n_1$ is unbiased for \bar{Y}_1 , but has a bias equal to $W_2 (\bar{Y}_1 - \bar{Y}_2)$ in estimating \bar{Y} .

2.1 Subsampling the Nonrespondents

Hansen and Hurwitz (1946) suggest drawing a subsample of size $m = n_2/k$, $k \geq 1$, from the n_2 nonrespondents and assume that responses are available from all of them. The sample mean $\bar{y}_{2m} = (\sum_1^m y_i)/m$ is unbiased for the mean \bar{y}_2 of the n_2 units. The estimator for \bar{Y} suggested by the above authors is

$$\hat{Y}_{HH} = w_1 \bar{y}_1 + w_2 \bar{y}_{2m}, \quad (2.1)$$

where $w_1 = (n_1/n)$ and $w_2 = (n_2/n)$.

For a given set of n_1 respondents and n_2 nonrespondents, this estimator is unbiased for $\bar{y} = w_1 \bar{y}_1 + w_2 \bar{y}_2 = (\sum_1^n y_i)/n$. Thus, it is unbiased for \bar{Y} .

The variance of this estimator is

$$V(\hat{Y}_{HH}) = \frac{(1-f)}{n} S^2 + W_2 \frac{(k-1)}{n} S_{2m}^2, \quad (2.2)$$

where $f = (n/N)$; see Cochran (1977, p. 371).

Let $s_1^2 = \sum_1^{n_1} (y_i - \bar{y}_1)^2 / (n_1 - 1)$ and $s_{2m}^2 = \sum_1^m (y_i - \bar{y}_{2m})^2 / (m - 1)$ denote the variances of the n_1 responses and the m subsampled units. An unbiased estimator of the variance is

$$\begin{aligned} v(\hat{Y}_{HH}) &= \frac{(1-f)}{n} \left[\frac{(n_1 - 1)s_1^2 + (n_2 - k)s_{2m}^2}{n - 1} \right] \\ &+ \frac{(1-f)}{n} \left[\frac{n_1 (\bar{y}_1 - \hat{Y}_{HH})^2 + n_2 (\bar{y}_{2m} - \hat{Y}_{HH})^2}{n - 1} \right] \\ &+ \frac{(N-1)w_2(k-1)s_{2m}^2}{N(n-1)}. \end{aligned} \quad (2.3)$$

This expression can also be obtained from the variance estimators for double sampling and stratification derived by Cochran (1977, p. 333) and Rao (1973); see also Rao (1983).

Post-stratification and subsampling

The $(n - n_1)$ nonrespondents may be classified into $(L - 1)$ strata of sizes (n_2, n_3, \dots, n_L) according to an auxiliary characteristic, or for convenience in sampling at the next phase. Subsamples of size $m_h = (n_h/k_h)$, $k_h \geq 1$, provide the means $\bar{y}_{hm} = \sum_1^{m_h} y_{hi}/m_h$ and variances $s_{hm}^2 = \sum_1^{m_h} (y_{hi} - \bar{y}_{hm})^2 / (m_h - 1)$.

The unbiased estimator for \bar{Y} now is

$$\hat{Y} = \sum_1^L w_h \bar{y}_{hm}, \quad (2.4)$$

where $w_h = (n_h/n)$ and $\bar{y}_{1m} = \bar{y}_1$.

The variance of the above estimator is

$$V(\hat{Y}) = \frac{(1-f)}{n} S^2 + \sum_2^L \frac{W_h(k_h-1)}{n} S_h^2 \quad (2.6)$$

where $S_h^2 = \Sigma_1^{N_h} (y_{hi} - \hat{Y}_h)^2 / (N_h - 1)$. The estimator for the variance is

$$\begin{aligned} v(\hat{Y}) = & \frac{(1-f)}{n} \sum_1^L \frac{(n_h - k_h) s_{hm}^2}{(n-1)} + \frac{(1-f)}{n} \sum_1^L \frac{n_h (\bar{y}_{hm} - \hat{Y})^2}{(n-1)} \\ & + \frac{(N-1)}{N(n-1)} \sum_2^L w_h (k_h - 1) s_{hm}^2, \end{aligned} \quad (2.7)$$

where $k_h = 1$, $y_{1m} = \bar{y}_1$, and $s_{1m}^2 = s_1^2$ as defined earlier.

Other types of post-stratification may be considered. For instance, the n units, respondents as well as the nonrespondents, may be post-stratified into L strata according to an auxiliary variable. The h -th stratum will now have n_{h1} respondents ($\Sigma_1^L n_{h1} = n_1$) with mean \bar{y}_{h1} and n_{h2} nonrespondents ($\Sigma_1^L n_{h2} = n_2$). A subsample of size $m_{h2} = (n_{h2}/k_h)$ from the n_{h2} units will provide the mean \bar{y}_{h2m} . An unbiased estimator for the mean \bar{Y}_h of the h -th stratum now is

$$\hat{Y}_h = \frac{n_{h1}\bar{y}_{h1} + n_{h2}\bar{y}_{h2m}}{n_h} \quad (2.8)$$

where $n_h = (n_{h1} + n_{h2})$, and the unbiased estimator for \bar{Y} is

$$\hat{Y} = \sum_1^L \frac{n_h}{n} \hat{Y}_h = \sum_1^L \frac{n_{h1}\bar{y}_{h1} + n_{h2}\bar{y}_{h2m}}{n}. \quad (2.9)$$

The variance of this estimator and its estimate can be found as in the above case.

The estimator in (2.4) is preferable if there is much difference among the means of the response and nonresponse strata. The estimator in (2.9) should be preferred if the means of the respondents and nonrespondents differ in each stratum, and if there is much difference among the means of the strata.

Sarndal and Swensson (1985) consider unequal probabilities of selection at the first phase and subsampling the nonrespondents after post-stratification.

3. RATIO ESTIMATORS

Let x_i , $i = (1, 2, \dots, N)$, denote an auxiliary characteristic with population mean $\bar{X} = (\Sigma_1^N x_i) / N$. Let \bar{X}_1 and \bar{X}_2 denote the means of the response and nonresponse groups. Let $\bar{x} = (\Sigma_1^n x_i) / n$ denote the mean of all the n units. Let $\bar{x}_1 = (\Sigma_1^{n_1} x_i) / n_1$ and $\bar{x}_2 = (\Sigma_1^{n_2} x_i) / n_2$ denote the means of the n_1 responding units and the n_2 nonresponding units. Further, let $\bar{x}_{2m} = (\Sigma_1^m x_i) / m$ denote the mean of the $m = (n_2/k)$ subsampled units.

The population variances of x and y are denoted by S_x^2 and S_y^2 , and the population covariance by S_{xy} . The correlation coefficient is $\rho_{xy} = (S_{xy}/S_x S_y)$. The sample variances are denoted by s_x^2 and s_y^2 . As before, the subscripts 1 and 2 denote the response and nonresponse groups.

3.1 The Conventional Estimator for the Mean

The ratio estimator for \bar{Y} is

$$t_1 = \frac{\bar{y}^*}{\bar{x}^*} \bar{X} = r^* \bar{X} \quad (3.1)$$

where \bar{y}^* is the same as \bar{Y}_{HH} in (2.1), $\bar{x}^* = (w_1 \bar{x}_1 + w_2 \bar{x}_{2m})$, and $r^* = (\bar{y}^*/\bar{x}^*)$; see Cochran (1977, p. 374). Now,

$$t_1 - \bar{Y} = \frac{(\bar{y}^* - R\bar{x}^*)\bar{X}}{\bar{x}^*} \doteq (\bar{y}^* - R\bar{x}^*) \left(1 - \frac{\bar{x}^* - \bar{X}}{\bar{X}}\right) \quad (3.2)$$

where $R = (\bar{Y}/\bar{X})$. The approximation in (3.2) is obtained by expressing $(1/\bar{x}^*)$ in Taylor's series, and it is valid for large values of the sample sizes n and m . From (3.2) the bias of t_1 is

$$B_1 = E(t_1 - \bar{Y}) \doteq \frac{(1-f)}{n\bar{X}} (RS_x^2 - S_{xy}) + \frac{W_2(k-1)}{n\bar{X}} (RS_{x2}^2 - S_{xy2}). \quad (3.3)$$

The bias vanishes only if (a) the regression of y on x goes through the origin for both the response and nonresponse strata and (b) the slopes of both the regressions are equal to R . The first condition is needed for the ratio estimator to be the optimum estimator for \bar{Y} . For the second condition to be satisfied, $R_2 = (\bar{Y}_2/\bar{X}_2)$ should not differ much from $R_1 = (\bar{Y}_1/\bar{X}_1)$.

From (3.2), a large sample approximation to the Mean Square Error (MSE) of t_1 is

$$M_1 = E(t_1 - \bar{Y})^2 \doteq \frac{(1-f)}{n} S_d^2 + W_2 \frac{(k-1)}{n} S_{d2}^2 \quad (3.4)$$

$$= \frac{(1-f)}{n} \sum_1^2 \frac{(NW_h - 1)}{(N-1)} S_{dh}^2 + W_2 \frac{(k-1)}{n} S_{d2}^2 \quad (3.4a)$$

where $S_d^2 = \Sigma_1^N (y_i - Rx_i)^2 / (N-1)$ and $S_{dh}^2 = \Sigma_1^{N_h} (y_{hi} - Rx_{hi})^2 / (N_h - 1)$ for $h = 1, 2$. The expression in (3.4) is briefly indicated by Cochran (1977).

An estimator for this MSE is obtained by replacing S_d^2 in (3.4a) by $s_{d1}^2 = \Sigma_1^{n_1} (y_i - r^*x_i)^2 / (n_1 - 1)$, S_{d2}^2 by $s_{d2}^2 = \Sigma_1^m (y_i - r^*x_i)^2 / (m - 1)$ and W_h by w_h . It is possible to suggest alternative estimators for the above MSE.

3.2 An Alternative Estimator for the Mean

In some situations, there may not be any nonresponse on the auxiliary characteristic. Family size, years of education, years of employment, and the like, are the above type of auxiliary variables.

The subsample provides the means \bar{x}_{2m} and \bar{y}_{2m} . However, since $\bar{x} = (\sum_1^n x_i)/n$ is available, for \bar{Y} we may consider

$$t_2 = \frac{\bar{y}^*}{\bar{x}} \bar{X} = \frac{w_1 \bar{y}_1 + w_2 \bar{y}_{2m}}{\bar{x}} \bar{X}. \quad (3.5)$$

Since the expectation of \bar{y}^* conditional on the first sample is equal to \bar{y} , the bias in t_2 is the same as the one in $\hat{Y}_R = (\bar{y}/\bar{x}) \bar{X}$. We note that \hat{Y}_R is the ratio estimator for the case of complete response. This result can also be derived from the expression

$$t_2 - \bar{Y} = \frac{\bar{y} - R\bar{x}}{\bar{x}} \bar{X} + \frac{\bar{y}^* - \bar{y}}{\bar{x}} \bar{X}. \quad (3.6)$$

Since the conditional mean of \bar{y}^* is equal to \bar{y} , the bias of t_2 is

$$B_2 = E(t_2 - \bar{Y}) = \frac{(1-f)}{n\bar{X}} (RS_x^2 - S_{xy}). \quad (3.7)$$

If the regression of y on x for the entire population goes through the origin, the bias of t_2 in (3.7) vanishes. If the regression for the second stratum also goes through the origin, the bias of t_1 in (3.3) would be small only when $R_2 = (\bar{Y}_2/\bar{X}_2)$ is close to R .

From (3.6), the MSE of t_2 is

$$M_2 = E(t_2 - \bar{Y})^2 = \frac{(1-f)}{n} S_d^2 + \frac{W_2(k-1)}{n} S_{y2}^2 \quad (3.8)$$

$$= \frac{(1-f)}{n} \frac{\sum (NW_h - 1) S_{dh}^2}{N-1} + W_2 \frac{(k-1)}{n} S_{y2}^2. \quad (3.8a)$$

Note that $S_d^2 = S_y^2 + R^2 S_x^2 - 2RS_{xy}$. An estimator of this MSE is obtained by replacing S_{d1}^2 , S_{d2}^2 , S_{y2}^2 , and W_h by s_{d1}^2 , s_{d2}^2 , s_{y2}^2 , and w_h respectively, where

$$s_{d1}^2 = \sum_{i=1}^{n_1} (y_i - r^{**}x_i)^2 / (n_1 - 1),$$

$$s_{d2}^2 = \sum_{i=1}^m (y_i - r^{**}x_i)^2 / (m - 1),$$

$$s_{y2}^2 = \sum_{i=1}^m (y_i - \bar{y}_{2m})^2 / (m - 1).$$

In these expressions, $r^{**} = (\bar{y}^*/\bar{x})$.

Comparing the approximate expressions in (3.4) and (3.8), we find that when $R_1 = (\bar{Y}_1/\bar{X}_1)$ does not differ much from $R_2 = (\bar{Y}_2/\bar{X}_2)$, t_2 will have smaller MSE than t_1 provided the correlation ρ_2 in the nonresponse stratum is not too high. Secondly, if R_1 differs much from R_2 , t_2 may have smaller MSE than t_1 even when ρ_2 is high. The following Section contains further comparisons between these two estimators.

3.3 Further Comparisons

In this Section, we compare t_1 and t_2 through the linear model. For the two groups, we consider the models

$$y_{1i} = \alpha_1 + \beta x_i + e_{1i}, \quad i = (1, 2, \dots, N_1) \quad (3.9a)$$

and

$$y_{2i} = \alpha_2 + \beta x_i + e_{2i}, \quad i = (1, 2, \dots, N_2), \quad (3.9b)$$

with the following assumptions:

$$E(e_{1i} | x_i) = 0, \quad E(e_{1i} e_{1i'}) = 0, \quad V(e_{1i} | x_i) = v_1 x_i^\ell;$$

$$E(e_{2i} | x_i) = 0, \quad E(e_{2i} e_{2i'}) = 0, \quad V(e_{2i} | x_i) = v_2 x_i^\ell.$$

We note that $(i \neq i')$ and in practice ℓ may lie between zero and 2. Further e_{1i} and e_{2i} are assumed to be uncorrelated. Biases and MSE's of t_1 and t_2 are obtained in the Appendix with the assumption that the response group of size N_1 and the nonresponse group of size N_2 are samples from the super-populations represented by the above models.

Comparisons of the biases

Let I denote the observations from the first initial sample. Since $E[(1/\bar{x}^*) | I] \geq (1/\bar{x})$ and $E(1/\bar{x}) \geq (1/\bar{X})$, from (A.2) and (A.3) we find that both t_1 and t_2 overestimate \bar{Y} . Further the bias B_1 of t_1 is larger than the bias B_2 of t_2 . From (A.6) and (A.7),

$$B_1 - B_2 = \frac{\alpha_w W_2 (k - 1) S_{x2}^2}{n \bar{X}^2} \quad (3.10)$$

This difference in the biases increases with the size of the nonresponse stratum and decreases with an increase in the size of the subsample.

Comparison of the MSE's

From (A.9) and (A.20), the difference in the MSE's of t_1 and t_2 is

$$M_1 - M_2 = (A_1 - A_2) - C_2 + (D_1 - D_2). \quad (3.11)$$

From (A.10), (A.21), and (A.22),

$$(A_1 - A_2) - C_2 = [3V(\alpha_w) + \alpha_w^2 - \beta^2 \bar{X}^2] \frac{W_2 (k - 1)}{n \bar{x}^2} S_{x2}^2. \quad (3.12)$$

We note that

$$\begin{aligned} V(\alpha_w) &= \alpha_1^2 V(w_1) + \alpha_2^2 V(w_2) + 2\alpha_1 \alpha_2 \text{Cov}(w_1, w_2) \\ &= \frac{N - n}{(N - 1)n} (\alpha_1 - \alpha_2)^2 W_1 W_2. \end{aligned} \quad (3.13)$$

The difference in (3.12) becomes large as α_1 and α_2 differ much from each other. A sufficient condition for the right side of (3.12) to be nonnegative is that $\alpha_W > \beta \bar{X}$. Further analysis of this result shows that the above difference becomes large if $C_x = (S_x/\bar{X})$ becomes larger than $C_y = (S_y/\bar{Y})$ as the correlation $\rho_{xy} = (S_{xy}/S_x S_y)$ increases.

From (A.12) and (A.24),

$$\begin{aligned} D_1 - D_2 &= E\{[2(\delta - \delta^*) + 3(\delta^{*2} - \delta^2)]\bar{e}^{*2}\} \\ &\quad + 2E[\delta^* - \delta - \delta^{*2} + \delta^2] \bar{E} \bar{e}^*. \end{aligned} \quad (3.14)$$

We note that $(\delta^* - \delta) = (\bar{x}^* - \bar{x})/\bar{X} = w_2(\bar{x}_{2m} - \bar{x}_2)/\bar{X}$. Further, $E(\delta^* - \delta) = 0$.

When $\ell = 0$, from (3.14) and the results in (A.14) and (A.17), to $O(n^{-2})$,

$$\begin{aligned} D_1 - D_2 &= 3E[(\delta^{*2} - \delta^2)\bar{e}^{*2}] - 2E[(\delta^{*2} - \delta^2)\bar{E} \bar{e}^*] \\ &= \frac{3W_2(k-1)S_{x2}^2}{n^2\bar{X}^2} (W_1v_1 + kW_2v_2) - \frac{2W_2(k-1)S_{x2}^2}{Nn\bar{X}^2} (W_1v_1 + W_2v_2) \\ &= \{[2(1-f) + 1](W_1v_1 + W_2v_2) + 3(k-1)W_2v_2\} \frac{W_2(k-1)}{n^2\bar{X}^2} S_{x2}^2. \end{aligned} \quad (3.15)$$

This expression clearly is nonnegative.

When $\ell = 1$, from (3.14), (A.15) and (A.16), to $O(n^{-1})$

$$\begin{aligned} D_1 - D_2 &= 2E\left[(\delta - \delta^*) \frac{(w_1v_1\bar{x}_1 + w_2kv_2\bar{x}_{2m})}{n}\right] \\ &\quad + 2E\left[(\delta^* - \delta) \frac{(w_1v_1\bar{x}_1 + w_2v_2\bar{x}_{2m})}{N}\right]. \end{aligned} \quad (3.16)$$

Noting that $E[(\delta^* - \delta)\bar{x}_1|I] = 0$, from (3.16),

$$\begin{aligned} D_1 - D_2 &= -(2/n) E[kw_2^2\bar{x}_{2m}(\bar{x}_{2m} - \bar{x}_2)]v_2 + (2/N) E[w_2^2\bar{x}_{2m}(\bar{x}_{2m} - \bar{x}_2)]v_2 \\ &= -(2/n)kE[w_2^2V(\bar{x}_{2m}|I)]v_2 + (2/N) E[w_2^2V(\bar{x}_{2m}|I)]v_2 \\ &= -\frac{2(Nk - n)W_2(k-1)S_{x2}^2}{Nn^2\bar{X}^2} v_2. \end{aligned} \quad (3.17)$$

Thus, when $\ell = 1$, $D_2 > D_1$. However, the difference in (3.17) becomes negligible when n is large.

The above results suggest that when $\ell = 0$, t_1 has larger MSE than t_2 if α is larger than $\beta\bar{X}$. When $\ell = 1$, t_1 will have larger MSE than t_2 if α is considerably larger than $\beta\bar{X}$.

4. SEPARATE RATIO ESTIMATORS

4.1 The First Estimator

If (\bar{X}_1, \bar{X}_2) are known, the separate ratio estimator for \bar{Y} that can be suggested is

$$\hat{\bar{Y}}_S = w_1 r_1 \bar{X}_1 + w_2 r_2 \bar{X}_2, \quad (4.1)$$

where $r_1 = (\bar{y}_1/\bar{x}_1)$ and $r_2 = (\bar{y}_2/\bar{x}_2)$. However, (\bar{X}_1, \bar{X}_2) can be estimated by (\bar{x}_1, \bar{x}_2) and $(\bar{y}_{2m}/\bar{x}_{2m})$ is an estimator of r_2 . With these estimates, an estimator for \bar{Y} is

$$t_3 = w_1 \bar{y}_1 + w_2 \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2. \quad (4.2)$$

This estimator can be used if \bar{x}_2 is available but \bar{X} is not; however, it does not make use of \bar{x}_1 .

From (4.2)

$$t_3 - \bar{Y} = (\bar{y} - \bar{Y}) + w_2 (\bar{x}_2/\bar{x}_{2m}) (\bar{y}_{2m} - r_2 \bar{x}_{2m}). \quad (4.3)$$

If m is large, from (4.3) the bias in t_3 is

$$B_3 = E(t_3 - \bar{Y}) \doteq \frac{(k-1)}{n\bar{X}_2} (R_2 S_{x2}^2 - S_{xy2}). \quad (4.4)$$

The MSE of t_3 is

$$M_3 = E(t_3 - \bar{Y})^2 \doteq \frac{(1-f)}{n} S_y^2 + \frac{W_2(k-1)}{n} S_{r2d2}^2 \quad (4.5)$$

where $S_{r2d2}^2 = \Sigma_1^{N_2} (y_i - R_2 x_i)^2 / (N_2 - 1)$.

An estimator for this MSE is obtained by replacing the first term on the right of (4.5) by $v(\bar{y}) = (1-f)s_y^2/n$, S_{r2d2}^2 by $s_{r2d2}^2 = \Sigma_1^m (y_i - r_{2m} x_i)^2 / (m-1)$, where $r_{2m} = (\bar{y}_{2m}/\bar{x}_{2m})$, and W_2 by w_2 .

4.2 The Second Estimator

An estimator that utilizes \bar{X} and \bar{x} is

$$t_4 = t_3 \left(\frac{\bar{X}}{\bar{x}} \right) = \left(w_1 \bar{y}_1 + w_2 \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2 \right) \left(\frac{\bar{X}}{\bar{x}} \right). \quad (4.6)$$

It may be beneficial to consider this estimator since the conditional mean of t_3 for large m is equal to \bar{y} , and hence the conditional expectation of t_4 becomes equal to $(\bar{y}/\bar{x})\bar{X}$.

From (4.6),

$$t_4 - \bar{Y} = \left(\frac{\bar{y}}{\bar{x}} \bar{X} - \bar{Y} \right) + w_2 \left(\frac{\bar{x}_2}{\bar{x}_{2m}} \right) (\bar{y}_{2m} - r_2 \bar{x}_{2m}) \left(\frac{\bar{X}}{\bar{x}} \right). \quad (4.7)$$

If n and m are large, the bias of t_4 is

$$B_4 = E(t_4 - \bar{Y}) \doteq \frac{(1-f)}{n\bar{X}} (RS_x^2 - S_{xy}) + \frac{(k-1)}{n\bar{X}_2} (R_2 S_{x2}^2 - S_{xy2}). \quad (4.8)$$

The MSE of t_4 is

$$\begin{aligned} M_4 &= E(t_4 - \bar{Y})^2 = \frac{(1-f)}{n} S_d^2 + W_2 \frac{(k-1)}{n} S_{r2d2}^2 \\ &= \frac{(1-f)}{n} \frac{\sum (NW_h - 1) S_{dh}^2}{N-1} + \frac{W_2 (k-1)}{n} S_{r2d2}^2. \end{aligned} \quad (4.9)$$

An estimator of M_4 is obtained by replacing S_{d1}^2 , S_{d2}^2 , S_{r2d2}^2 , and W_2 by s_{d1}^2 , s_{d2}^2 , s_{r2d2}^2 , and w_2 respectively, where

$$\begin{aligned} s_{d1}^2 &= \sum_{i=1}^{n_1} (y_i - r_{x_i}^*)^2 / (n_1 - 1), \\ s_{d2}^2 &= \sum_{i=1}^m (y_i - r_{x_i}^*)^2 / (m - 1), \\ s_{r2d2}^2 &= \sum_{i=1}^m (y_i - r_{2m} x_i)^2 / (m - 1). \end{aligned}$$

We note that $r^* = (\bar{y}^* / \bar{x}^*)$ as defined in Section (3.1).

Comparing (4.5) and (4.9), we find that t_4 will have smaller MSE than t_3 if the population correlation between x and y is high.

Further investigation is needed to evaluate the merits of the above two separate estimators relative to the estimators in the previous Section.

5. RATIO ESTIMATORS IN THE PRESENCE OF THE HARDCORE

It is becoming increasingly apparent that in spite of subsampling the nonrespondents and a number of call-backs, a significant proportion of the sampled units, the hard-core, do not respond to the items in the survey.

For this situation, we consider the population to be composed of three groups of sizes (N_1, N_2, N_3) , $N = \sum_1^3 N_i$, with means $(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)$ and variances $(S_{y1}^2, S_{y2}^2, S_{y3}^2)$. The means and variances for the auxiliary characteristic are $(\bar{X}_1, \bar{X}_2, \bar{X}_3)$ and $(S_{x1}^2, S_{x2}^2, S_{x3}^2)$. The population means of these two items are $\bar{Y} = (W_1 \bar{Y}_1 + W_2 \bar{Y}_2 + W_3 \bar{Y}_3)$ and $\bar{X} = (W_1 \bar{X}_1 + W_2 \bar{X}_2 + W_3 \bar{X}_3)$, where $\sum_1^3 W_i = 1$. Let $R_1 = (\bar{Y}_1 / \bar{X}_1)$, $R_2 = (\bar{Y}_2 / \bar{X}_2)$ and $R_3 = (\bar{Y}_3 / \bar{X}_3)$.

In the initial sample of size n , only n_1 units respond and provide the means (\bar{x}_1, \bar{y}_1) . The number of units (n_2, n_3) in the last two groups are not known, but their sum $(n_2 + n_3) = (n - n_1)$ is known. The means (\bar{x}_2, \bar{x}_3) of the auxiliary characteristic may be known, but (\bar{y}_2, \bar{y}_3) for the item of interest are not observed.

We consider the situation where in the subsample of size $m = (n - n_1) / k$, only m_2 units respond and provide the means $(\bar{x}_{2m}, \bar{y}_{2m})$. The remaining $m_3 = (m - m_2)$ units, the "hard-core", do not respond. Note that m_1 is not defined.

In Rao and Jackson (1984), a number of estimators for \bar{Y} for the above situation are examined, without utilizing the auxiliary information. In this Section, we suggest the following six estimators that utilize the additional information. We briefly present the conditions for which these estimators may be the optimum ones. For the sake of space, we have not presented the derivations for these estimators.

- (I). The difference between R_1 , R_2 and R_3 is negligible. The m_3 units of the third group, the hard-core, is a random subsample of the m_2 respondents at the second phase. In this case,

$$\hat{Y}_{H1} = \frac{n_1 \bar{y}_1 + (n - n_1) \bar{y}_{2m}}{n_1 \bar{x}_1 + (n - n_1) \bar{x}_{2m}} \bar{X}. \quad (5.1)$$

- (II). Same conditions as in I, but poor correlation in the second and third groups. For this case,

$$\hat{Y}_{H2} = \frac{n_1 \bar{y}_1 + (n - n_1) \bar{y}_{2m}}{n \bar{x}} \bar{X}. \quad (5.2)$$

- (III). $\bar{X}_3 = (N_1 \bar{X}_1 + N_2 \bar{X}_2) / (N_1 + N_2)$ and $\bar{Y}_3 = (N_1 \bar{Y}_1 + N_2 \bar{Y}_2) / (N_1 + N_2)$, and (R_1, R_2, R_3) do not differ much from each other. Under these conditions,

$$\hat{Y}_{H3} = \frac{n_1 \bar{y}_1 + km_2 \bar{y}_{2m}}{n_1 \bar{x}_1 + km_2 \bar{x}_{2m}} \bar{X}. \quad (5.3)$$

Note that, since $E(m_2/m) = n_2 / (n - n_1)$, an unbiased estimator of n_2 is $[(n - n_1) / m] m_2 = km_2$.

- (IV). Same conditions as in (III), but poor correlation in the second and third groups. For this case,

$$\hat{Y}_{H4} = \frac{n_1 \bar{y}_1 + km_2 \bar{y}_{2m}}{(n_1 + km_2) \bar{x}} \bar{X}. \quad (5.4)$$

- (V). The three ratios differ from one another. The n_3 units of the third group are a random subsample from the n_2 units of the second group. In this case,

$$\hat{Y}_{H5} = \left[\frac{n_1}{n} \bar{y}_1 + \frac{(n - n_1)}{n} \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2 \right] \left(\frac{\bar{X}}{\bar{x}} \right). \quad (5.5)$$

- (VI). The three ratios differ from one another. The n_3 units of the third group are a random subsample from the $(n_1 + n_2)$ units of the first two groups. Under these conditions,

$$\hat{Y}_{H6} = \left(\frac{n_1}{n_1 + km_2} \bar{y}_1 + \frac{km_2}{n_1 + km_2} \frac{\bar{y}_{2m}}{\bar{x}_{2m}} \bar{x}_2 \right) \left(\frac{\bar{X}}{\bar{x}} \right). \quad (5.6)$$

While we expect the above conditions to be satisfactory, further research is needed to evaluate the performances of the above six estimators.

ACKNOWLEDGMENTS

The author would like to thank Dr. J.N.K. Rao and Dr. M.P. Singh for their interest in this topic. Thanks also to the referee for making suggestions towards improving the presentation of the material.

APPENDIX: BIASES AND MSE'S UNDER THE SUPER POPULATION MODEL

Let $\alpha_w = W_1\alpha_1 + W_2\alpha_2$, $\alpha_w = w_1\alpha_1 + w_2\alpha_2$,

$$\bar{E} = \sum_1^N e_i/N, \bar{e}_1 = \sum_1^{n_1} e_i/n_1, \bar{e}_{2m} = \sum_1^m e_i/m \text{ and } \bar{e}^* = w_1\bar{e}_1 + w_2\bar{e}_{2m}.$$

Now

$$\bar{Y} = \alpha_w + \beta\bar{X} + \bar{E}, \quad (\text{A.1})$$

$$t_1 - \bar{Y} = \frac{\bar{X}}{\bar{X}^*}\alpha_w - \alpha_w + \frac{\bar{e}^*}{\bar{X}^*}\bar{X} - \bar{E}, \quad (\text{A.2})$$

and

$$t_2 - \bar{Y} = \frac{\bar{X}}{\bar{X}}\alpha_w - \alpha_w + \beta\left(\frac{\bar{X}^*}{\bar{X}} - 1\right)\bar{X} + \frac{\bar{e}^*}{\bar{X}}\bar{X} - \bar{E}. \quad (\text{A.3})$$

1. Biases

Let $\delta^* = (\bar{X}^* - \bar{X})/\bar{X}$ and $\delta = (\bar{x} - \bar{X})/\bar{X}$. Taylor's expansion about \bar{X} gives

$$\frac{\bar{X}}{\bar{X}^*} = 1 - \delta^* + \delta^{*2} \dots \quad (\text{A.4})$$

and

$$\frac{\bar{X}}{\bar{x}} = 1 - \delta + \delta^2 \dots \quad (\text{A.5})$$

With these expansions, from (A.2) and (A.3), to $O(n^{-1})$ the biases of t_1 and t_2 are

$$B_1 = \frac{V(\bar{X}^*)}{\bar{X}^2}\alpha_w = \left[\frac{(1-f)}{n\bar{X}^2}S_x^2 + \frac{W_2(k-1)}{n\bar{X}^2}S_{x2}^2 \right]\alpha_w \quad (\text{A.6})$$

and

$$B_2 = \frac{V(\bar{x})}{\bar{X}^2}\alpha_w = \left[\frac{(1-f)}{n\bar{X}^2}S_x^2 \right]\alpha_w. \quad (\text{A.7})$$

2. Mean Square Error of t_1

From the expansion in (A.4),

$$\left(\frac{\bar{X}}{\bar{X}^*}\right)^2 = 1 - 2\delta^* + 3\delta^{*2}. \quad (\text{A.8})$$

From (A.2), the MSE of t_1 can be written as

$$M_1 = E(t_1 - \bar{Y})^2 = A_1 + D_1, \quad (\text{A.9})$$

where

$$\begin{aligned}
 A_1 &= E \left(\frac{\bar{X}}{\bar{X}^*} \alpha_w - \alpha_w \right)^2 \\
 &\doteq E \left[(1 - 2\delta^* + 3\delta^{*2}) \alpha_w^2 \right] + \alpha_w^2 - 2E \left[(1 - \delta^* + \delta^{*2}) \alpha_w \right] \\
 &= V(\alpha_w) + \left[3E(\alpha_w^2) - 2\alpha_w^2 \right] \left[V(\bar{X}^*) / \bar{X}^2 \right] \\
 &= V(\alpha_w) + \left[3V(\alpha_w) + \alpha_w^2 \right] \left[V(\bar{X}^*) / \bar{X}^2 \right] \quad (\text{A.10})
 \end{aligned}$$

and

$$D_1 = E \left(\frac{\bar{e}^*}{\bar{X}^*} \bar{X} - \bar{E} \right)^2. \quad (\text{A.11})$$

With the expansions in (A.4) and (A.8)

$$\begin{aligned}
 \left(\frac{\bar{e}^*}{\bar{X}^*} \bar{X} - \bar{E} \right)^2 &= \left(\frac{\bar{X}}{\bar{X}^*} \right)^2 \bar{e}^{*2} + \bar{E}^2 - 2 \left(\frac{\bar{X}}{\bar{X}^*} \right) \bar{E} \bar{e}^* \\
 &\doteq (1 - 2\delta^* + 3\delta^{*2}) \bar{e}^{*2} + \bar{E}^2 - 2(1 - \delta^* + \delta^{*2}) \bar{E} \bar{e}^* \\
 &= (\bar{e}^* - \bar{E})^2 - (2\delta^* - 3\delta^{*2}) \bar{e}^{*2} + 2(\delta^* - \delta^{*2}) \bar{E} \bar{e}^*. \quad (\text{A.12})
 \end{aligned}$$

Now,

$$\bar{e}^{*2} = w_1^2 \bar{e}_1^2 + w_2^2 \bar{e}_{2m}^2 + 2w_1 w_2 \bar{e}_1 \bar{e}_{2m}. \quad (\text{A.13})$$

Thus, conditional on n_1 and n_2 , when $\ell = 0$,

$$E(\bar{e}^{*2} | n_1, n_2) = \frac{w_1^2}{n_1} v_1 + \frac{k w_2^2}{n_2} v_2 = \frac{w_1 v_1 + k w_2 v_2}{n}. \quad (\text{A.14})$$

Similarly, when $\ell = 1$,

$$\begin{aligned}
 E(\bar{e}^{*2} | n_1, n_2) &= \frac{w_1^2}{n_1^2} v_1 \left(\sum_1^{n_1} x_i \right) + \frac{(k w_2)^2}{n_2^2} v_2 \left(\sum_1^m x_i \right) \\
 &= \frac{1}{n} (w_1 v_1 \bar{x}_1 + w_2 k v_2 \bar{x}_{2m}). \quad (\text{A.15})
 \end{aligned}$$

Further,

$$\bar{E} \bar{e}^* = \frac{1}{N} \left[\sum_1^{n_1} e_i + \sum_1^m e_i + \sum_1^{N-(n_1+m)} e_i \right] (w_1 \bar{e}_1 + w_2 \bar{e}_{2m}) \quad (\text{A.16})$$

From (A.16), when $t = 0$,

$$E(\bar{E}\bar{e}^*) = \frac{1}{N}(w_1 v_1 + w_2 v_2). \quad (\text{A.17})$$

Similarly, when $\ell = 1$,

$$E(\bar{E}\bar{e}^*) = \frac{1}{N}(w_1 v_1 \bar{x}_1 + w_2 v_2 \bar{x}_{2m}). \quad (\text{A.18})$$

3. Mean Square Error of t_2

From the expansion in (A.5)

$$\left(\frac{\bar{X}}{\bar{x}}\right)^2 \doteq 1 - 2\delta + 3\delta^2. \quad (\text{A.19})$$

From (A.10), the MSE of t_2 can be written as

$$M_2 = E(t_2 - \bar{Y})^2 = A_2 + C_2 + D_2. \quad (\text{A.20})$$

With the expansions in (A.5) and (A.19)

$$\begin{aligned} A_2 &= E\left(\frac{\bar{X}}{\bar{x}}\alpha_w - \alpha_w\right)^2 \\ &\doteq E\left[(1 - 2\delta + 3\delta^2)\alpha_w^2\right] + \alpha_w^2 - 2E\left[(1 - \delta + \delta^2)\alpha_w\right]\alpha_w \\ &= V(\alpha_w) + \left[3E(\alpha_w^2) - 2\alpha_w^2\right]\left[V(\bar{x})/\bar{x}^2\right] \\ &= V(\alpha_w) + \left[3V(\alpha_w) + \alpha_w^2\right]\left[V(\bar{x})/\bar{x}^2\right], \end{aligned} \quad (\text{A.21})$$

$$\begin{aligned} C_2 &= \beta^2 E\left(\frac{\bar{x}^* - \bar{x}}{\bar{x}}\right)^2 \bar{x}^2 \\ &\doteq \beta^2 E(\bar{x}^* - \bar{x})^2 = \beta^2 E\left[w_2^2(\bar{x}_{2m} - \bar{x}_2)^2\right] \\ &= \beta^2 W_2 \frac{(k-1)}{n} S_{x2}^2, \end{aligned} \quad (\text{A.22})$$

and

$$D_2 = E\left(\frac{\bar{e}^*}{\bar{x}}\bar{X} - \bar{E}\right)^2. \quad (\text{A.23})$$

With the expansions in (A.5) and (A.19)

$$\left(\frac{\bar{X}}{\bar{x}}\bar{e}^* - \bar{E}\right)^2 = (\bar{e}^* - \bar{E})^2 - (2\delta - 3\delta^2)\bar{e}^{*2} + 2(\delta - \delta^2)\bar{E}\bar{e}^*. \quad (\text{A.24})$$

We note that

$$E\left[\frac{\bar{X}}{\bar{x}}(\alpha_w - \alpha_w)\left(\frac{\bar{x}^* - \bar{x}}{\bar{x}}\right)\bar{X}\right] \doteq E[(\bar{x}^* - \bar{x})(\alpha_w - \alpha_w)] = 0. \quad (\text{A.25})$$

REFERENCES

- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.
- HANSEN, M.H., and HURWITZ, W.N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.
- RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- RAO, P.S.R.S. (1983). Randomization approach. In *Incomplete Data in Sample Surveys*, Vol. 2; (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 33-44.
- RAO, P.S.R.S., and JACKSON, J.E. (1984). Estimation through the procedure of subsampling the nonrespondents. Presented at the American Statistical Association Meetings, Philadelphia.
- SÄRNDAL, C.E., and SWENSSON, B. (1985). Incorporating nonresponse modelling in a general randomization theory approach. *Proceedings of the 45th Session of the International Statistical Institute*, Section 15.2.