Research Journal of Applied Sciences 2 (6): 769-772, 2007

Estimation of the Population Mean Using Difference Cum Ratio Estimator With Full Response on the Auxiliary Character

A.A. Sodipo and K.O Obisesan
Department of Statistics, University of Ibadan, Ibadan, Nigeria

Abstract: A number of researchers have investigated the problem of nonresponse in sample surveys and the effect of its bias on the estimates obtained. Some of these researchers include: Hansen and Hurwitz, Hendricks, Deming, Houseman, Durbin and Stuart, Kish and Hess, Ericson, Kokan, Sodipo, Okafor and Lee, among others. Rao suggested a ratio estimator for the population mean $\bar{Y}$, of a study variable y in the presence of nonresponse, when the population mean $\bar{X}$ of an auxiliary character x with full response (full data available on both the respondents and nonrespondents) is known. Our present research proposes a difference cum ratio estimator which is essentially an extension of Rao ratio estimator. The proposed estimator proves to be more efficient than and also contains Rao ratio estimator. We study the properties of our proposed estimator and discuss procedures for determining the optimum values of the sample size n and the subsampling rate k. An illustration is made with a hypothetical data to demonstrate the practical applicability of our new estimator.

Key words: Difference cum ratio, estimator, mean, auxiliary character, nonresponse

INTRODUCTION

Practical situations do exist in which, aside our knowledge of $\bar{X}$, we may be fortunate also to have full response (information available for all the respondents and nonrespondents) on the auxiliary variable x, even though there is nonresponse on the study variate y. For instance, in a practical situation in which we consider it desirable to estimate the average household expenditure on food in a particular month for a specified human population, our auxiliary variable x can be the household size. We may cheaply collect this auxiliary information during the listing period, rather than during the survey period. The collection of the auxiliary information during the listing period has the advantage of affording a good knowledge of $\bar{X}$, if the latter has been hitherto unknown. This is because the household size may be obtained directly from the potential respondents and indirectly (from neighbours) from the potential nonrespondents. However, if $\bar{X}$ is known, the auxiliary information may be collected from the sample units during the survey rather than from the whole population during the listing period. This full response on the auxiliary character x, together with the known $\bar{X}$, are then utilized in constructing our proposed difference cum ratio estimator of the population mean $\bar{Y}$.

Rao (1986) suggested a ratio estimator based on the full response on the auxiliary variable x, whose population mean $\bar{X}$ is known. His proposed estimator is

$$e_1 = \frac{\bar{y}^*}{\bar{x}}\bar{X} = \hat{R}^*\bar{X} \quad (1)$$

Where

$$\bar{y}^* = w_1\bar{y}_1 + w_2\bar{y}_{2m}, \quad \hat{R}^* = \frac{\bar{y}^*}{\bar{x}}$$

and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Rao gave its large sample bias, M.S.E. and an estimator of this M.S.E., respectively as

$$B(e_1) \doteq \frac{(1-f)}{n\bar{X}}\left(RS_x^2 - S_{yx}\right) \quad (2)$$

$$M(e_1) \doteq \frac{(1-f)}{n}S_{d,R}^2 + \frac{W_2(k-1)}{n}S_{y_2}^2 \quad (3)$$

Where $S_{d,R}^2 = S_y^2 - 2RS_{yx} + R^2S_x^2$

and

$$\hat{M}(e_1) \doteq \frac{(1-f)}{n}\sum_{h=1}^{2}\frac{(Nw_h - 1)}{(N-1)}s_{dhm,\hat{R}^*}^2 + \frac{w_2(k-1)}{n}s_{y_{2m}}^2 \quad (4)$$

Where $s_{d1m,\hat{R}^*}^2 = s_{d1,\hat{R}^*}^2 = \frac{1}{(n_1-1)}\sum_{i=1}^{n_1}\left(y_i - \hat{R}^*x_i\right)^2$

Corresponding Author: A.A. Sodipo, Department of Statistics, University of Ibadan, Ibadan, Nigeria

$$s^2_{d2m,\hat{R}^*} = \frac{1}{(m-1)}\sum_{i=1}^{m}\left(y_i - \hat{R}^* x_i\right)^2$$

$$s^2_{y_{2m}} = \frac{1}{(m-1)}\sum_{i=1}^{m}\left(y_i - \bar{y}_{2m}\right)^2 \text{ and } \hat{R}^* = \frac{\bar{y}^*}{\bar{x}} \text{ as before.}$$

Remark 1: Although Rao (1986) did not give the optimum values of n and k for his ratio estimator e_1, however, using the cost function with the expected cost C_0 given by

$$C_0 = c_0 n + c_1 n W_1 + \frac{c_2 n W_2}{k} \tag{5}$$

Where c_0 is the unit cost of initial sampling of the n units, c_1 is the unit cost of processing returns from the respondents and c_2 is the unit cost of collecting and processing data from the stratum of nonrespondents, we obtain

$$k_0 = \left[\frac{c_2\left(S^2_{d,R} - W_2 S^2_{y_2}\right)}{(c_0 + c_1 W_1) S^2_{y_2}}\right]^{\frac{1}{2}} \tag{6}$$

and

$$n_0 = \frac{N\left[S^2_{d,R} - W_2(k-1)S^2_{y_2}\right]}{\left[NM_0 + S^2_{d,R}\right]} \tag{7}$$

Where M_0 is the fixed value of $M(e_1)$ in (3).

SAMPLING DESIGN

Consider a finite population consisting of N units from which we randomly select without replacement an initial sample of fixed size n and information collected on the study variate y on which there is nonresponse and also on the auxiliary character x with complete response. Suppose n_1 of the n units in the sample respond to the survey variable y, while $n_2 = n-n_1$ units do not. We note here however, that all the n sample units (respondents and nonrespondents) supply the auxiliary information x, on which there is no nonresponse. We then apply Hansen and Hurwitz (1946) scheme for subsampling the nonrespondents and select $m = n_2/k$, $k \geq 1$ (fixed in advance), from whom we collect the required information.

Now, let E_2, V_2 and C_2 denote conditional expectation, variance and covariance operators respectively, taken over all possible subsamples of the nonrespondents. E_1, V_1 and C_1 denote unconditional operators over all possible samples.

We note that

$$E(\bar{y}^*) = E_{12}(\bar{y}^*) = E_1(\bar{y}) = \bar{Y}$$

$$E(\bar{x}) = E_1(\bar{x}) = \bar{X}, \quad V(\bar{x}) = V_1(\bar{x}) = \frac{(1-f)}{n}S^2_x$$

$$E(w_h) = E_1(w_h) = W_h, \qquad h = 1,2$$

Also $C_2(\bar{y}^*, \bar{x}) = 0$, $C_1(\bar{y}, \bar{x}) = \frac{(1-f)}{n}S_{yx}$

and hence $C(\bar{y}^*, \bar{x}) = \frac{(1-f)}{n}S_{yx}$.

THE PROPOSED DIFFERENCE CUM RATIO ESTIMATOR WITH ITS BIAS, MEAN SQUARE ERROR AND OPTIMUM VALUES OF k AND n

We now consider our proposed difference cum ratio estimator together with its properties.

Estimator:

$$e_2 = \left[\bar{y}^* - \theta(\bar{x} - \bar{X})\right]\left(\frac{\bar{X}}{\bar{x}}\right) \tag{8}$$

Where θ is a suitably chosen constant.

Bias:

$$B(e_2) \doteq \frac{(1-f)}{n\bar{X}}\left[(\theta + R)S^2_x - S_{yx}\right] \tag{9}$$

Mean square error:

$$M(e_2) \doteq \frac{(1-f)}{n}S^2_{d,(\theta+R)} + \frac{W_2(k-1)}{n}S^2_{y_2} \tag{10}$$

Where

$$S^2_{d,(\theta+R)} = S^2_y - 2(\theta + R)S_{yx} + (\theta + R)^2 S^2_x$$

Estimator of M.S.E:

$$\hat{M}(e_2) \doteq \hat{V}(\bar{y}^*) - 2\left(\theta + \hat{R}^*\right)\hat{C}(\bar{y}^*, \bar{x}) + \left(\theta + \hat{R}^*\right)^2 \hat{V}(\bar{x}) \tag{11}$$

where $\hat{V}(\bar{y}^*) = \frac{(1-f)}{n}\left[\frac{(n_1-1)\,s^2_{y_1} + (n_2-k)\,s^2_{y_{2m}}}{(n-1)}\right] +$

$$\frac{(1-f)}{n}\left[\frac{n_1(\bar{y}_1 - \bar{y}^*)^2 + n_2(\bar{y}_{2m} - \bar{y}^*)^2}{(n-1)}\right] + \frac{(N-1)w_2(k-1)\,s^2_{y_{2m}}}{N(n-1)}$$

$$\hat{C}\left(\bar{y}^{*},\bar{x}\right)=\frac{\left(1-f\right)}{n}\left[\frac{\left(n_1-1\right)s_{yx_1}+\left(n_2-k\right)s_{yx_{2m}}}{\left(n-1\right)}\right]+\frac{\left(1-f\right)}{n}\left[\frac{n_1\left(\bar{y}_1-\bar{y}^{*}\right)\left(\bar{x}_1-\bar{x}^{*}\right)+n_2\left(\bar{y}_{2m}-\bar{y}^{*}\right)\left(\bar{x}_{2m}-\bar{x}^{*}\right)}{\left(n-1\right)}\right]+\frac{\left(N-1\right)w_2\left(k-1\right)s_{yx_{2m}}}{N\left(n-1\right)}$$

and

$$\hat{V}\left(\bar{x}\right)=\frac{\left(1-f\right)}{n}\left[\frac{\left(n_1-1\right)s_{x_1}^2+\left(n_2-k\right)s_{x_{2m}}^2}{\left(n-1\right)}\right]+\frac{\left(1-f\right)}{n}\left[\frac{n_1\left(\bar{x}_1-\bar{x}^{*}\right)^2+n_2\left(\bar{x}_{2m}-\bar{x}^{*}\right)^2}{\left(n-1\right)}\right]+\frac{\left(N-1\right)w_2\left(k-1\right)s_{x_{2m}}^2}{N\left(n-1\right)}$$

Optimum k and n:

$$k^{*}=\left[\frac{c_2\left(S_{d,(\theta+R)}^2-W_2\,S_{y_2}^2\right)}{\left(c_0+c_1\,W_1\right)S_{y_2}^2}\right]^{\frac{1}{2}} \tag{12}$$

and

$$n^{*}=\frac{N\left[S_{d,(\theta+R)}^2-W_2\left(k-1\right)S_{y_2}^2\right]}{\left[NM^{*}+S_{d,(\theta+R)}^2\right]} \tag{13}$$

Where M^{*} is the fixed value of $M(e_2)$.

By making use of the derivations, let ρ^{*} and β be defined as stated below:

$$\rho^{*}=\frac{C\left(\bar{y}^{*},\bar{x}\right)}{\sqrt{V\left(\bar{y}^{*}\right)}\sqrt{V\left(\bar{x}\right)}}=\frac{\frac{\left(1-f\right)}{n}S_{yx}}{\left[\frac{\left(1-f\right)}{n}S_y^2+\frac{W_2\left(k-1\right)}{n}S_{y_2}^2\right]^{\frac{1}{2}}\cdot\left[\frac{\left(1-f\right)}{n}S_x^2\right]^{\frac{1}{2}}}=\frac{\left(1-f\right)S_{yx}}{\left[\left(1-f\right)S_y^2+W_2\left(k-1\right)S_{y_2}^2\right]^{\frac{1}{2}}\cdot\left[\left(1-f\right)S_x^2\right]^{\frac{1}{2}}} \tag{14}$$

and

$$\beta=\frac{C\left(\bar{y}^{*},\bar{x}\right)}{V\left(\bar{x}\right)}=\frac{\frac{\left(1-f\right)}{n}S_{yx}}{\frac{\left(1-f\right)}{n}S_x^2}=\frac{S_{yx}}{S_x^2} \tag{15}$$

Remark 2: The value of θ that minimizes $M(e_2)$ in (10) is

$$\theta_o=\beta-R \tag{16}$$

which results in a minimum value of $M(e_2)$ given as

$$M_o\left(e_2\right)=\left(1-\rho^{*2}\right)\left[\frac{\left(1-f\right)}{n}S_y^2+\frac{W_2\left(k-1\right)}{n}S_{y_2}^2\right] \tag{17}$$

since (14) and (15) together become

$$\beta=\rho^{*}\frac{\left[\frac{\left(1-f\right)}{n}S_y^2+\frac{W_2\left(k-1\right)}{n}S_{y_2}^2\right]^{\frac{1}{2}}}{\left[\frac{\left(1-f\right)}{n}S_x^2\right]^{\frac{1}{2}}}=\rho^{*}\frac{\left[\left(1-f\right)S_y^2+W_2\left(k-1\right)S_{y_2}^2\right]^{\frac{1}{2}}}{\left[\left(1-f\right)S_x^2\right]^{\frac{1}{2}}} \tag{18}$$

Remark 3: If we substitute $\theta=\theta_o=\beta$-R in (9), $B(e_2)=0$. This shows that the optimum value of θ, that is $\theta_o=\beta$-R, minimizes the M.S.E. of our proposed difference cum ratio estimator and also makes it unbiased up to the first order of approximation.

Remark 4: We interestingly observe that if we insert $\theta=0$ into Eq. 8-13, respectively, they pairwisely and successively become Eq. 1-6. This means that when we choose the value of $\theta = 0$ in our estimator and its properties, $e_2 = e_1$. In other words, our proposed difference cum ratio estimator e_2 contains Rao (1986) ratio estimator e_1.

Remark 5: In order that e_2 be more efficient than $\bar{y}^{*}$, we choose θ such that $-R<\theta<2\beta-R$ in those cases where β and R have the same sign; while θ should be chosen such that $2\beta-R<\theta<-R$ in cases where β and R have opposite signs.

PRACTICAL INVESTIGATION

The Table 1 shows the percentage gain in efficiency of our difference cum ratio estimator e_2, over Hansen and Hurwitz (1946) estimator $\bar{y}^{*}$ and Rao (1986) ratio

Table 1: Percentage gain in efficiencies of estimators

Estimator		Estimated mean	Estimated M.S.E.	stimated efficiency	Percentage gain
$\bar{y}^*$		204.45	126.10	1.00	-
e_1		186.49	79.51	1.59	59
e_2	$\theta = -3.16$	195.97	52.23	2.41	141
	$\theta = \theta_o = -2.16$**	192.96**	44.76**	2.82**	182**
	$\theta = -1.16$	189.96	52.23	2.41	141

** Optimum values

estimator e_1 using hypothetical data with N = 100, n = 40, n_1 = 30, n_2 = 10, k = 2, m = 5, β = 3.301517, R = 5.459279 and ρ^* = 0.803111.

Remark 6: We see from the table that our proposed difference cum ratio estimator e_2 is unbiased (Remark 3) and more than three times as efficient as Rao (1986) ratio estimator e_1, which is biased.

Remark 7: We observe that all our practical results confirm our theoretical results.

REFERENCES

Deming, W.E., 1953. On the probability mechanism to attain an economic balance between the resulting error of response and bias of nonresponse. J. Am. Stat. Assoc., 48: 743-772

Durbin, J. and A. Stuart, 1954. Callbacks and clustering in sample surveys: An experimental study. J. Royal Stat. Soc., 117: 387-428.

Ericson, W.A., 1967. Optimal sample design with nonresponse. J. Am. Stat. Assoc., 62: 63-78.

Hansen, M.H. and W.N. Hurwitz, 1946. The problem of nonresponse in sample surveys. J. Am. Stat. Assoc., 41: 517-529.

Hendricks, W.A., 1949. Adjustment for bias by nonresponse in mailed surveys. Agric. Econ. Res., 1: 52-56.

Houseman, E.E., 1953. Statistical treatment of the nonresponse problem. Agric. Econ. Res., 5: 12-18.

Kish, L. and I. Hess, 1959. A 'replacement' procedure for reducing the bias of nonresponse. Am. Stat., 13: 17.19.

Kokan, A.R., 1986. On a model of nonresponse. J. Nig. Stat. Assoc., 3: 53-66.

Okafor, F.C. and H. Lee, 2000. Double sampling for ratio and regression estimation with subsampling the nonrespondents. Survey Methodol., 26: 183-188.

Rao, P.S.R.S., 1986. Ratio estimation with subsampling the nonrespondents. Survey Methodol., 12: 217-230.

Sodipo, A.A., 1997. Ratio estimator for the population mean in a double (Two-phase) sampling scheme with nonresponse. J. Sci. Res., 3: 1-3.

Sodipo, A.A., 1998. A ratio-type estimator for the population mean with subsampling and poststratification of the nonrespondents. J. Nig. Stat., pp: 44-53.

Sodipo, A.A., 2003a. Estimation of the population mean using difference and regression estimators in a double sampling design with nonresponse. Abacus J. Mathe. Assoc. Nig., 30: 214-220.

Sodipo, A.A., 2003b. Classes of estimators for the population mean of a variable of interest with nonresponse. Abacus J. Mathe. Assoc. Nig., 30: 304-311.