Mathematical Theory and Modeling www.iiste.org
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) DOI: 10.7176/MTM
Vol.9, No.7, 2019

Modified Ratio Estimator of Finite Population Total in Stratified Random Sampling Under Non-Response

David Oyoo[1*], Moses Manene[2], Christopher Ouma[3] and George Muhua[4]

[1,2,4] School of Mathematics, University of Nairobi, P. O. Box 30197-00100,Nairobi-Kenya
[3] School of Pure and Applied Sciences, Kenyatta University, P. O. Box 43844-00100, Nairobi-Kenya
***Corresponding Author's E-mail address:** koyoo2010@gmail.com

Abstract

In this paper, we propose a separate ratio-type estimator of finite population total. The procedure of subsampling non-respondents suggested by Hansen and Hurwitz (1946) is considered. Asymptotic properties of the proposed estimator are studied under stratified random sampling. Study of the asymptotic properties shows that the suggested estimator is unbiased and consistent. We recommend that simulation study should be done to check for performance of the estimator.

Key Words: *Stratified Random Sampling, Ratio Estimation, Non-Response*
DOI: 10.7176/MTM/9-7-08 Publication date: July 31st, 2019

1 Introduction

In survey sampling, estimation of finite population total with or without auxiliary information under complete data has been considered by several authors. Many authors have utilized some known population parameters of the auxiliary variable to suggest various ratio-type estimators. The suggested estimators have, however, focused on estimation of population total using complete data. Thus, as a gap to be filled, we focus on estimation of finite population total in this paper.

2 Non-Response in Sample Surveys

Non-response in sample surveys occurs when there is a failure to measure or to make observation on some units in the selected sample (Cochran, 1977). Non-response divides study population into two disjoint 'strata', where the first stratum consist of population units for which measurements would be obtained if the units are samples while the other stratum consist of population units for which measurements would not be obtained (Cochran, 1977). Various methods of correcting non-response have been extensively discussed in literature (Oyoo and Ouma, 2014). Such methods include imputation, resampling, partial deletion, weight adjustment and Hansen and Hurwitz subsampling method. In this paper, we use the Hansen and Hurwitz (1946) method since the method increases the weighted response rates of non-responses while reducing sampling costs is sub-sampling (Cochran, 1977). We shall use X and Y to denote the auxiliary variable and response variable respectively.

3 Hansen-Hurwitz Method (1946)

In this method, we begin by determining a sample size required to attain desired level of precision. We can let N_1 and N_2 be the respective number of population units in the responding and non-responding 'strata'with the corresponding sample sizes n_1 and n_2. Also, we let the respective proportion of response and non-response groups be $W_1 = \frac{N_1}{N}$ and $W_2 = \frac{N_2}{N}$, with the corresponding sample and population means as $\overline{y}_1, \overline{Y}_1$ and $\overline{y}_2, \overline{Y}_2$. Hansen and Hurwitz (1946) suggested that from the n_2 non-respondents, we draw a sub-sample of size $m = \frac{n_2}{h}, h \geq 1$. We assume that the sub-sample has a complete data so that the sample mean pair for the auxiliary variable and the study variable can be denoted as $(\overline{x}_{2m}, \overline{y}_{2m})$. However, in case of an auxiliary variable, the sub-sampled mean for the auxiliary variable can be denoted as $\overline{x}_2 m$. Using a single variable, Hansen and Hurwitz (1946) suggests an estimator for $\bar{Y}$ as

$$\overline{Y}_{HH} = w_1\overline{y}_1 + w_2\overline{y}_{2m} \tag{1}$$

Where $w_1 = \frac{n_1}{n}, w_2 = \frac{n_2}{n}$ and $n_1 + n_2 = n$

The estimator given in Eq. (1) is unbiased for $\overline{y} = w_1\overline{y}_1 + w_2\overline{y}_2$, which is further unbiased for

$$\overline{y}^* = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

Furthermore, the estimator given in Eq. (2) is unbiased for $\overline{Y}$. Therefore, based on the Hansen-Hurwitz Method, the unbiased estimator for the finite population total is

$$\widehat{Y}_{HH} = N(w_1\overline{y}_1 + w_2\overline{y}_{2m}) \tag{3}$$

The estimator given in (3) is obtained by applying the weight adjustment technique indirectly. Hansen and Hurwitz (1946) method has been studied and applied by various studies such as studies by Rao (1986), Walsh (1970), Reddy (1973), (1967) and Khoshnevisan et al (2007), who constructed a general family of estimators for the population mean using known values of some population parameters as

$$\hat{\overline{Y}}_k = \overline{y}\{\frac{a\overline{X}}{\alpha(a\overline{x} + b) + (1-\alpha)(a\overline{X} + b)}\}^g \tag{4}$$

where $a(\neq 0)$, b are either real numbers or functions of known parameters of the auxiliary variable.

This family of estimators was not, however, constructed under non-response. Under non-response, Kumar (2012) utilized known population parameters to construct a general family of estimators of population mean as

$$\hat{\overline{Y}} = \overline{y}^*(\frac{a\overline{x}^* + b}{a\overline{X} + b})^{\alpha}(\frac{a\overline{x} + b}{a\overline{X} + b})^{\beta} \tag{5}$$

where $a(\neq 0)$, b are either real numbers or functions of known parameters of the auxiliary variable and (α, β) are suitable chosen constants. By varying the values of the constants, various estimators have been constructed.

Few studies such as studies by Saghir and Shabbir (2012), Chaudhary et al. (2013) and Singh and Malik (2014) among others, have used Hansen-Hurwitz method under non-response to estimate finite population mean using auxiliary variable. However, despite extensive application of Hansen-Hurwitz method, most of these previously constructed ratio-type estimators have produced biased results. It is this gap of constructing biased ratio-type estimators under non-response that this study fills by constructing an unbiased ratio-type estimator for finite population total in stratified random sampling scheme and using the Hansen-Hurwitz method to take care of missing values.

The usual ratio estimator of finite population total Y_T in stratified random sampling using k strata is

$$\hat{Y}_T = \sum_{c=1}^{k} N_c r_c \overline{X}_c, c = 1, 2, ..., k \tag{6}$$

where $N_c, \overline{X}_c, r_c = \frac{\overline{y}_c}{\overline{x}_c}$ are the population size, population mean of the auxiliary variable and the usual ratio estimate respectively in stratum c

And $Bias(\hat{Y}_T)$ in Eq. (6) is given by

$$Bias(\hat{Y}_T) = \sum_{c=1}^{k} N_c \overline{Y}_c (\frac{N_c - n_c}{n_c N_c})\{C_{Xc}^2 - \rho_c C_{Xc} C_{Yc}\} \tag{7}$$

Where C_{Xc} and C_{Yc} are the respective population coefficients of variation of X and Y in stratum c and ρ_c is population correlation coefficient between X and Y X and Y in stratum c.

Under non-response, we use Hansen-Hurwitz method so that

$$\hat{Y}_T = \sum_{c=1}^{k} N_c r_c^* \overline{X}_c, c = 1, 2, ..., k \tag{8}$$

where r^* is obtained using the $\overline{y}$ as expressed in Eq. (1) not as expressed in Eq. (6).

Cochran (1977) and Daroga and Chaudhary (2002) define an unbiased ratio-type of population mean under simple random sampling scheme as

$$\hat{\overline{Y}} = \overline{r}\overline{X} + \frac{n(N-1)}{N(n-1)}(\overline{y} - \overline{r}\overline{x}) \tag{9}$$

where $\overline{r} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

In this paper, we use the expressions in Eqs. (1), (8) and (9) to construct a ratio-type estimator of finite population total stratified random sampling.

4 Notations and Definition of Symbols

We shall let Y_{cij} be the i^{th} population unit in group j in stratum c, where $i = 1, 2, ..., N$, $c = 1, 2, ..., k$ and $j = 1, 2$ (such that $j = 1-$ responding and $j = 2-$non-responding groups). Also, we shall let the overall population total Y_T be $Y_T = \sum_{c=1}^{k}\sum_{j=1}^{2}\sum_{i=1}^{N_{cj}} Y_{cij} = \sum_{c=1}^{k}\sum_{j=1}^{2} Y_{Tcj} = \sum_{c=1}^{k} Y_{Tc}$ where N_{cj} is the population size in stratum c in the j^{th} group so that N_c is the population size in stratum c and that $N = \sum_{c=1}^{k} N_c = \sum_{c=1}^{k}\sum_{j=1}^{2} N_{cj}$ is the overall population size. Further, we shall have $\overline{Y}_{cj} = \frac{1}{N_{cj}}\sum_{i=1}^{N_{cj}} Y_{cij}$ as the population mean for the j^{th} subgroup in stratum c, $\overline{Y}_c = \frac{1}{N_c}\sum_{j=1}^{2} N_{cj}\overline{Y}_{cj}$ as the population mean for

ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) DOI: 10.7176/MTM
Vol.9, No.7, 2019

stratum c and $\overline{Y} = \frac{1}{N}\sum_{c=1}^{k} N_c\overline{Y}_c$ as the overall population mean.

We shall also define the usual stratum population ratio as $R_c = \frac{\overline{Y}_c}{\overline{X}_c}$ and the i^{th} observation ratio in stratum c for j^th group as $R_{cij} = \frac{Y_{cij}}{X_{cij}}$ and that $R_{cj} = \sum_{i=1}^{N_{cj}} \frac{Y_{cij}}{X_{cij}}$ so that $\overline{R}_{cj} = \frac{1}{N_{cj}}\sum_{i=1}^{N_{cj}} \frac{Y_{cij}}{X_{cij}}$ and $\overline{R}_c = \frac{1}{N_c}\sum_{i=1}^{2} R_{cj}$. For variances, we shall let $S^2_{Yc} = \frac{1}{N_c-1}\sum_{i=1}^{N_c}(Y_{ci} - \overline{Y}_c)^2$ and $S^2_{Xc} = \frac{1}{N_c-1}\sum_{i=1}^{N_c}(X_{ci} - \overline{X}_c)^2$ to be the respective adjusted population variances for Y and X in stratum c. For co-variances, we shall use $S_{XYcj} = \frac{1}{N_{cj}-1}\sum_{i=1}^{N_{cj}}(X_{cij} - \overline{X}_{cj})(Y_{cij} - \overline{Y}_{cj})$ and $S_{RXcj} = \frac{1}{N_{cj}-1}\sum_{i=1}^{N_{cj}}(R_{cij} - \overline{R}_{cj})(X_{cij} - \overline{X}_{cj})$ to denote population co-variances between X and Y and between R and X in stratum c for group j.

For sample statistics, we shall use the corresponding lower cases so that we shall have n_{cj} as the sample size in stratum c in the j^{th} group so that $n_c = n_{c1} + n_{c2}$ is the sample size in stratum c and $n = \sum_{c=1}^{k} n_c$ is the overall population size , while m_c is the sub-sample size from the non-responding n_{c2} units. For the totals and means, the corresponding lower cases for the sample shall be y_{tcj}, y_{tc}, y_t and $\overline{y}_{cj}$, $\overline{y}_c$, $\overline{y}$ respectively. But under non-response, we shall define y_{tcj} as $y_{tc1} = n_{c1}\overline{y}_{c1}$ and $y_{tc2} = n_{c2}\overline{y}_{c2m}$.

Similar expressions shall apply for the auxiliary variable X

The corresponding sample ratios shall be $r_c = \frac{\overline{y}_c}{\overline{x}_c}$, $r_{cij} = \frac{y_{cij}}{x_{cij}}$, $r_{cj} = \sum_{i=1}^{n_{cj}} \frac{y_{cij}}{x_{cij}}$, $\overline{r}_{cj} = \frac{1}{n_{cj}}\sum_{i=1}^{n_{cj}} \frac{y_{cij}}{x_{cij}}$ and $\overline{r}_c = \frac{1}{n_c}\sum_{i=1}^{2} r_{cj}$. The corresponding sample variances and co-variances shall be expressed using lower cases of the population variances and co-variances above.

5 Proposed Estimator

Under non-response in the study variable and using the Hansen-Hurwitz Method, we suggest a ratio-type estimator of finite population total, denoted by Y_D, in stratified random sampling as

$$Y_D = \sum_{c=1}^{k}\sum_{j=1}^{2}[\overline{r}_{cj}X_{Tcj} + \frac{N_{cj}-1}{n_{cj}-1}(y_{tcj} - \overline{r}_{cj}x_{tcj})] \tag{10}$$

6 Derivation of the Proposed Estimator

Consider the two population groups of responding and non-responding population units with sizes N_1 and N_2 and that the corresponding population totals are Y_{T1} and Y_{T2} such that

$$Y_T = \sum_{c=1}^{k} Y_{Tc} = \sum_{c=1}^{k}[Y_{Tc1} + Y_{Tc2}]$$

That is,

$$Y_T = \sum_{c=1}^{k}[\overline{R}_{c1}X_{Tc1} + \overline{R}_{c2}X_{Tc2}] \tag{11}$$

Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) DOI: 10.7176/MTM
Vol.9, No.7, 2019

In the subsequent steps we shall consider a particular stratum c (say) to do the derivations and proofs.

Under non-response, the usual ratio estimator for the finite population total in stratum c is given by

$$\widehat{Y}_{Tc} = \bar{r}_{c1} X_{Tc1} + \bar{r}_{c2} X_{Tc2}$$

where $\bar{r}_{c1}$ and $\bar{r}_{c2}$ are as previously defined.

But we know that, $Bias(\widehat{Y}_{Tc}) = E(\widehat{Y}_{Tc}) - Y_{Tc}$, which can be expanded as,

$$Bias(\widehat{Y}_{Tc}) = [X_{Tc1} E(\bar{r}_{c1}) - Y_{Tc1}] + [X_{Tc2} E(\bar{r}_{c2}) - Y_{Tc2}]$$

That is, $Bias(\widehat{Y}_{Tc}) = Bias(Y_{Tc1}) + Bias(Y_{Tc2})$

But under Simple Random Sampling Without Replacement (SRSWOR),

$Cov(\bar{x}_{c1}, \bar{y}_{c1}) = \frac{N_c - n_c}{n_c N_c} S_{XYc1}$, $Cov(\bar{x}_{c2m}, \bar{y}_{c2m}) = \frac{n_{c2} - m_c}{m_c n_{c2}} S_{XYc2}$
and
$Cov(\bar{r}_{c1}, \bar{x}_{c1}) = \frac{N_c - n_c}{n_c N_c} S_{RXc1}$, $Cov(\bar{r}_{c2}, \bar{x}_{c2m}) = \frac{n_{c2} - m_c}{m_c n_{c2}} S_{RXc2}$

where,

But, $S_{RXcj} = \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} (R_{cij} - \overline{R}_{cj})(X_{cij} - \overline{X}_{cj})$, for $j = 1, 2$

Now, S_{RXcj} can further be expanded as follows,

$$S_{RXcj} = \frac{1}{N_{cj}-1} [\sum_{i=1}^{N_{cj}} \frac{Y_{cij}}{X_{cij}} X_{cij} - N_{cj} \overline{R}_{cj} \overline{X}_{cj}] = \frac{1}{N_{cj}-1} [Y_{Tcj} - \overline{R}_{cj} X_{Tcj}]$$

That is,

$$S_{RXcj} = \frac{1}{N_{cj}-1} [Y_{Tcj} - X_{Tcj} E(\overline{R}_{cj})] = -\frac{1}{N_{cj}-1} Bias(\hat{Y}_{Tcj})$$

Therefore,

$Cov(\bar{r}_{c1}, \bar{x}_{c1}) = -\frac{N_{c1} - n_{c1}}{n_{c1} N_{c1}} \frac{1}{N_{c1}-1} Bias(\hat{Y}_{Tc1})$ so that,

$$Bias(\widehat{Y}_{Tc1}) = -\frac{n_{c1} N_{c1} (N_{c1}-1)}{N_{c1} - n_{c1}} Cov(\bar{r}_{c1}, \bar{x}_{c1})$$

That is, $Bias(\widehat{Y}_{Tc1}) = -\frac{n_c N_c (N_{c1}-1)}{N_c - n_c} \frac{N_c - n_c}{n_c N_c} S_{RXc1}$

which reduces to, $Bias(\widehat{Y}_{Tc1}) = -(N_{c1} - 1) S_{RXc1}$.

Therefore, the estimator of the Bias of $\widehat{Y}_{Tc1})$ is given by $\widehat{Bias(\widehat{Y}_{Tc1})} = -(N_{c1}-1)s_{rxc1}$, where,

$s_{rxc1} = \frac{1}{n_{c1}-1}\sum_{i=1}^{n_{c1}}(r_{ci1}-\overline{r}_{c1})(x_{ci1}-\overline{x}_{c1}) = \frac{1}{n_{c1}-1}[\sum_{i=1}^{n_{c1}} r_{ci1} - n_{c1}\overline{r}_{c1}\overline{x}_{c1}]$

which can further be expressed as,

$s_{rxc1} = \frac{1}{n_{c1}-1}[\sum_{i=1}^{n_{c1}} \frac{y_{ci1}}{x_{ci1}}x_{ci1} - n_{c1}\overline{r}_{c1}\overline{x}_{c1}] = \frac{1}{n_{c1}-1}[\sum_{i=1}^{n_{c1}} y_{ci1} - n_{c1}\overline{r}_{c1}\overline{x}_{c1}]$

That is, $s_{rxc1} = \frac{1}{n_{c1}-1}[y_{tc1} - \overline{r}_{c1}x_{tc1}]$.

Thus, the Bias in $\widehat{Y}_{Tc1}$ reduces to,

$Bias(\widehat{Y}_{Tc1}) = -\frac{(N_{c1}-1)}{(n_{c1}-1)}[y_{tc1} - \overline{r}_{c1}x_{tc1}]$

Now,

$E[\widehat{Y}_{Tc1} - Bias(\widehat{Y}_{Tc1})] = Y_{Tc1}$, which implies that,

$$Y_{Tc1} = [\widehat{Y}_{Tc1} + \frac{(N_{c1}-1)}{(n_{c1}-1)}[y_{tc1} - \overline{r}_{c1}x_{tc1}] \quad (12)$$

Similarly, using the same procedure for non-responding group, we obtain

$Bias(\widehat{Y}_{Tc2}) = -(n_{c2}-1)s_{rxc2}$

so that, $Bias(\widehat{Y}_{Tc2}) = -\frac{(n_{c2}-1)}{(m_c-1)}[y_{tc2} - \overline{r}_{c2}x_{tc2}]$

Assuming proportional allocation of sample sizes in the responding groups such that $\frac{N_{c1}-1}{n_{c1}-1} \approx \frac{N_{c2}-1}{nc2-1}$ and that there is a high response rate in the second sampling phase such that m_c, is so close to n_{c2}, then we can write

$Bias(\widehat{Y}_{Tc2}) = -\frac{N_{c2}-1}{n_{c2}-1}[y_{tc2} - \overline{r}_{c2}x_{tc2}]$

So that,

$$Y_{Tc2} = [\widehat{Y}_{Tc2} + \frac{(N_{c2}-1)}{(n_{c2}-1)}[y_{tc2} - \overline{r}_{c2}x_{tc2}]] \quad (13)$$

But we know that $\widehat{Y}_{Tcj} = \overline{r}_{cj}X_{Tcj}$ so that using Eq. (10) and Eq. (13), we obtain,

$$Y_{Tc} = \sum_{j=1}^{2}[\overline{r}_{cj}X_{Tcj} + \frac{(N_{cj}-1)}{(n_{cj}-1)}[y_{tcj} - \overline{r}_{cj}x_{tcj}]]$$

Summing over the entire population, we obtain

$$\widehat{Y}_T = \sum_{c=1}^{k}\sum_{j=1}^{2}[\overline{r}_{cj}X_{Tcj} + \frac{N_{cj}-1}{n_{cj}-1}(y_{tcj} - \overline{r}_{cj}x_{tcj})] = Y_D$$

Hence the derivation.

7 Asymptotic Properties of the Proposed Estimator

7.1 Unbiasedness

In this section, we shall show that Y_D is an unbiased estimator of the population total Y_T. We first consider the following lemma.

Lemma 7.1 *The sample ratio mean for the j^{th} group in stratum c, $\overline{r}_{cj}$, is unbiased for the population ratio mean for the j^{th} group in stratum c, $\overline{R}_{cj}$*

Proof. To prove the lemma, we need to show that $E(\overline{r}_{cj}) = \overline{R}_{cj}$.

Now, $E(\overline{r}_{cj}) = E(\frac{1}{n_{cj}}\sum_{i=1}^{n_{cj}} R_{cij})$

That is, $E(\overline{r}_{cj}) = \frac{1}{n_{cj}}\sum_{i=1}^{n_{cj}}\frac{1}{N_{cj}}\sum_{i=1}^{N_{cj}}\frac{Y_{cij}}{X_{ci}}$

which implies that,

$$E(\overline{r}_{cj}) = \frac{1}{n_{cj}}\sum_{i=1}^{n_{cj}}\overline{R}_{cj} = \frac{1}{n_{cj}}n_{cj}\overline{R}_{cj} = \overline{R}_{cj}$$

Hence the proof.

□

Similarly, it can also be shown that the sample ratio mean in stratum c, $\overline{r}_c$ is unbiased for population ratio mean in stratum c, $\overline{R}_c$

Theorem 7.1. *The estimator Y_D is unbiased estimator of the finite population total Y_T under the assumption that the response in the second phase sample is such that m_c, is so close to n_{c2}*

Proof. In this proof, we need to show that $E(Y_D) = Y_T$.

That is, we need to show that $E(Y_D) = E[\sum_{c=1}^{k}\sum_{j=1}^{2}[\overline{r}_{cj}X_{Tcj} + \frac{N_{cj}-1}{n_{cj}-1}(y_{tcj} - \overline{r}_{cj}x_{tcj})] = Y_T$

Now, $E(Y_D)$ can be expanded as shown below

$$E(Y_D) = \sum_{c=1}^{k}[E(\overline{r}_{c1}X_{Tc1} + \overline{r}_{c2}X_{Tc2}) + \frac{N_{c1}-1}{n_{c1}-1}E(y_{tc1} - \overline{r}_{c1}x_{tc1}) + \frac{N_{c2}-1}{n_{c2}-1}E(y_{tc2} - \overline{r}_{c2}x_{tc2})] \quad (14)$$

But from Lemma 7.1, we have, $E(\overline{r}_{c1}X_{Tc1} + \overline{r}_{c2}X_{Tc2}) = \overline{R}_{c1}X_{Tc1} + \overline{R}_{c2}X_{Tc2}$ so that $E(Y_D)$ becomes

$$E(Y_D) = \sum_{c=1}^{k}[(\overline{R}_{c1}X_{Tc1} + \overline{R}_{c2}X_{Tc2}) + \frac{N_{c1}-1}{n_{c1}-1}E(y_{tc1} - \overline{r}_{c1}x_{tc1}) + \frac{N_{c2}-1}{n_{c2}-1}E(y_{tc2} - \overline{r}_{c2}x_{tc2})] \quad (15)$$

Now,

$$\begin{aligned} y_{tc1} - \overline{r}_{c1}x_{tc1} &= n_{c1}\overline{y}_{c1} - n_{c1}\overline{r}_{c1}\overline{x}_{c1} \\ &= \sum_{i=1}^{n_{c1}} y_{ci1} - n_{c1}\overline{r}_{c1}\overline{x}_{c1} \\ &= \sum_{i=1}^{n_{c1}} \frac{y_{ci1}}{x_{ci1}}x_{ci1} - n_{c1}\overline{r}_{c1}\bar{x}_{c1} \\ &= \sum_{i=1}^{n_{c1}}(x_{ci1} - \bar{x}_{c1})(r_{ci1} - \overline{r}_{c1}) \end{aligned}$$

which reduces to,

$$y_{tc1} - \overline{r}_{c1}x_{tc1} = (n_{c1} - 1)s_{rxc1} \quad (16)$$

Similarly,

$$y_{tc2} - \overline{r}_{c2}x_{tc2} = n_{c2}\overline{y}_{c2} - n_{c2}\overline{r}_{c2}\overline{x}_{c2}$$

That is

$$y_{tc1} - \overline{r}_{c1}x_{tc1} = n_{c2}\sum_{i=1}^{m_c}\frac{1}{m_c}y_{ci2} - m_c\overline{r}_{c2}\overline{x}_{c2} \quad (17)$$

But assuming that m_c is large and is close to n_{c2} such that $m_c \approx n_{c2}$, $\forall c = 1, 2, ...k$, then Eq. (17) reduces to,

$$y_{tc1} - \overline{r}_{c1}x_{tc1} = (m_c - 1)s_{rxc2} \quad (18)$$

Therefore, substituting Eq. (16) and Eq. (18) in Eq. (15) and assuming that $m_c \approx n_{c2}$, we obtain,

$$E(Y_D) = \sum_{c=1}^{k}[(\overline{R}_{c1}X_{Tc1} + \overline{R}_{c2}X_{Tc2}) + \frac{(N_{c1}-1)}{(n_{c1}-1)}E(n_{c1}-1)(s_{rxc1}) + \frac{(N_{c2}-1)}{(n_{c2}-1)}E(m_c-1)(s_{rxc2})]$$

But for $m_c \approx n_{c2}$, $\forall c = 1, 2, ...k$, $E(Y_D)$ simplifies to

$$E(Y_D) = \sum_{c=1}^{k}[(\overline{R}_{c1}X_{Tc1} + \overline{R}_{c2}X_{Tc2}) + (N_{c1}-1)E(s_{rxc1}) + (N_{c2}-1)E(s_{rxc2})] \quad (19)$$

That is,

$$E(Y_D) = \sum_{c=1}^{k}[(\overline{R}_{c1}X_{Tc1} + \overline{R}_{c2}X_{Tc2}) + (N_{c1}-1)S_{rxc1} + (N_{c2}-1)S_{rxc2}] \tag{20}$$

But,

$S_{XYcj} = \frac{1}{N_{cj}-1}\sum_{i=1}^{N_{cj}}(X_{cij} - \overline{X}_{cj})(Y_{cij} - \overline{Y}_{cj}) = \frac{1}{N_{cj}-1}(Y_{Tcj} - N_{cj}\overline{R}_{cj}\overline{X}_{cj})$ so that Eq, (20) becomes

$$E(Y_D) = \sum_{c=1}^{k}[(\overline{R}_{c1}X_{Tc1} + \overline{R}_{c2}X_{Tc2}) + (Y_{Tc1} - N_{c1}\overline{R}_{c1}\overline{X}_{c1}) + (Y_{Tc2} - N_{c2}\overline{R}_{c2}\overline{X}_{c2})]$$

That is,

$$E(Y_D) = \sum_{c=1}^{k}[Y_{Tc1} + Y_{Tc2}] = \sum_{c=1}^{k} Y_{Tc} = Y_T$$

Hence the proof. □

7.2 Mean Squared Error (MSE)

By definition, MSE of the estimator Y_D for the population total Y_T is obtained as follows,

$MSE(Y_D) = E[Y_D - Y_T]^2$,

which can be expressed as $MSE(Y_D) = E[Y_D + E(Y_D) - E(Y_D) - Y_T]^2$

That is, $MSE(Y_D) = E[Y_D - E(Y_D)]^2 + [E(Y_D) - Y_T]^2$, which reduces to,

$$MSE(Y_D) = Var(Y_D) + [Bias(Y_D)]^2 \tag{21}$$

But under the assumption that $m_c \approx n_{c2}$, $Bias(Y_D) = 0$, so that Eq. (21) reduces to,

$$MSE(Y_D) = Var(Y_D) \tag{22}$$

7.3 Variance of Y_D

In this section, we show how variance of the suggested estimator is obtained. By definition, variance of a random variable X is obtained as $Var(X) = E(X^2) - [E(X)]^2$ and we proceed as shown below;

Theorem 7.2. *Under the assumption of a large sample size, variance of the suggested estimator Y_D is given as*

$$Var(Y_D) = \sum_{c=1}^{k}[(X_{Tc1} - \frac{n_{c1}(N_{c1}-1)}{(n_{c1}-1)}\overline{x}_{c1})^2\frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}}S^2_{Rc1} + (X_{Tc2} - \frac{n_{c2}(N_{c2}-1)}{(n_{c2}-1)}\overline{x}_{c2m})^2\frac{N_{c2}-m_c}{m_cN_{c2}}S^2_{Rc2}] \tag{23}$$

Proof. In this proof, we wish to show that $Var(Y_D)$ is as given in Eq. (23). From Eq. (10), we recall that,

$$Y_D = \sum_{c=1}^{k}\sum_{j=1}^{2}[\overline{r}_{cj}X_{Tcj} + \frac{N_{cj}-1}{n_{cj}-1}(y_{tcj} - \overline{r}_{cj}x_{tcj})]$$

which can be rearranged and expanded as,

$$Y_D = \sum_{c=1}^{k}[(X_{Tc1} - \tfrac{n_{c1}(N_{c1}-1)}{(n_{c1}-1)}\overline{x}_{c1})\overline{r}_{c1} + (X_{Tc2} - \tfrac{n_{c2}(N_{c2}-1)}{(n_{c2}-1)}\overline{x}_{c2m})\overline{r}_{c2} + \tfrac{n_{c1}(N_{c1}-1)}{n_{c1}-1}\overline{y}_{c1} + \tfrac{n_{c2}(N_{c2}-1)}{n_{c2}-1}\overline{y}_{c2m}]$$

That is,

$$Y_D = \sum_{c=1}^{k}[A\overline{r}_{c1} + B\overline{r}_{c2} + C] \tag{24}$$

where,

$A = (X_{Tc1} - \frac{n_{c1}(N_{c1}-1)}{(n_{c1}-1)}\overline{x}_{c1})$, $B = (X_{Tc2} - \frac{n_{c2}(N_{c2}-1)}{(n_{c2}-1)}\overline{x}_{c2m})$ and $C = \frac{n_{c1}(N_{c1}-1)}{n_{c1}-1}\overline{y}_{c1} + \frac{n_{c2}(N_{c2}-1)}{n_{c2}-1}\overline{y}_{c2m}$ and A, B, C are all constants that depend only on sample and population sizes, totals and means.

But we know that for any two random variables X and Y and constants a, b and c,

$$Var(aX + bY + c) = a^2Var(X) + a^2Var(X) + 2abCov(X,Y)$$

so that from Eq. (24), we have;

$$Var(Y_D) = \sum_{c=1}^{k}[A^2Var(\overline{r}_{c1}) + B^2Var(\overline{r}_{c2})] \tag{25}$$

since $Cov(\overline{r}_{c1}, \overline{r}_{c2}) = 0$

But under SRSWOR, $Var(\overline{r}_{c1}) = \frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}}S^2_{Rc1}$ and $Var(\overline{r}_{c2}) = \frac{N_{c2}-m_c}{m_cN_{c2}}S^2_{Rc2}$, where $S^2_{Rcj} = \frac{1}{N_{cj}-1}\sum_{i=1}^{N_{cj}}(R_{cij} - \overline{R}_{cj})^2$

Therefore, substituting in Eq. (25), we obtain;

$$Var(Y_D) = \sum_{c=1}^{k}[A^2\frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}}S^2_{Rc1} + B^2\frac{N_{c2}-m_c}{m_cN_{c2}}S^2_{Rc2}] \tag{26}$$

Replacing the expressions for A and B in Eq. (26) we get,

$$Var(Y_D) = \sum_{c=1}^{k}[(X_{Tc1} - \tfrac{n_{c1}(N_{c1}-1)}{(n_{c1}-1)}\overline{x}_{c1})^2\tfrac{N_{c1}-n_{c1}}{n_{c1}N_{c1}}S^2_{Rc1} + (X_{Tc2} - \tfrac{n_{c2}(N_{c2}-1)}{(n_{c2}-1)}\overline{x}_{c2m})^2\tfrac{N_{c2}-m_c}{m_cN_{c2}}S^2_{Rc2}]$$

Hence the proof. □

Therefore, the unbiased estimator of $Var(Y_D)$ becomes

Mathematical Theory and Modeling www.iiste.org
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) DOI: 10.7176/MTM
Vol.9, No.7, 2019

$Var(Y_D) = \sum_{c=1}^{k}[(X_{Tc1} - \frac{n_{c1}(N_{c1}-1)}{(n_{c1}-1)}\overline{x}_{c1})^2 \frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}} s^2_{rc1} + (X_{Tc2} - \frac{n_{c2}(N_{c2}-1)}{(n_{c2}-1)}\overline{x}_{c2m})^2 \frac{N_{c2}-m_c}{m_c N_{c2}} s^2_{rc2}]$

Where $s^2_{rc1} = \frac{1}{n_{c1}-1}\sum_{i=1}^{n_{c1}}(r_{ci1} - \overline{r}_{c1})^2$ and $s^2_{rc2} = \frac{1}{m_c-1}\sum_{i=1}^{m_c}(r_{ci2} - \overline{r}_{c2})^2$ are the unbiased estimators of S^2_{Rc1} and S^2_{Rc2} respectively.

Corollary 1: For sufficiently large sample size, the unbiased estimator of we have

$$Var(Y_D) = \sum_{c=1}^{k}[\overline{x}^2_{c1}\frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}}s^2_{rc1} + \overline{x}^2_{c2m}\frac{N_{c2}-m_c}{m_c N_{c2}}s^2_{rc2}] \tag{27}$$

Proof. We have shown that

$$Var(Y_D) = \sum_{c=1}^{k}[(X_{Tc1} - \frac{n_{c1}(N_{c1}-1)}{(n_{c1}-1)}\overline{x}_{c1})^2 \frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}} S^2_{Rc1} + (X_{Tc2} - \frac{n_{c2}(N_{c2}-1)}{(n_{c2}-1)}\overline{x}_{c2m})^2 \frac{N_{c2}-m_c}{m_c N_{c2}} S^2_{Rc2}]$$

Now, for a sufficiently large sample size, we have $(n_{cj} - 1) \approx n_{cj}$ so that,

$$Var(Y_D) = \sum_{c=1}^{k}[(X_{Tc1} - (N_{c1}-1)\bar{x}_{c1})^2 \frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}} s^2_{rc1} + (X_{Tc2} - (N_{c2}-1)\bar{x}_{c2m})^2 \frac{N_{c2}-m_c}{m_c N_{c2}} s^2_{rc2}]$$

That is,

$$Var(Y_D) = \sum_{c=1}^{k}[(X_{Tc1} - N_{c1}\overline{x}_{c1} + \overline{x}_{c1})^2 \frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}} s^2_{rc1} + (X_{Tc2} - N_{c2}\overline{x}_{c2m} + \overline{x}_{c2m})^2 \frac{N_{c2}-m_c}{m_c N_{c2}} s^2_{rc2}]$$

But assuming that $\overline{x}_{c1}$ and $\overline{x}_{c2m}$ are close estimators of $\overline{X}_{c1}$ and $\overline{X}_{c2}$ respectively so that $N_{c1}\overline{x}_{c1} \approx X_{Tc1}$ and $N_{c2}\overline{x}_{c2m} \approx X_{Tc2}$, then $Var(Y_D)$ becomes,

$$Var(Y_D) = \sum_{c=1}^{k}[(X_{Tc1} - X_{Tc1} + \overline{x}_{c1})^2 \frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}} s^2_{rc1} + (X_{Tc2} - X_{Tc2} + \overline{x}_{c2m})^2 \frac{N_{c2}-m_c}{m_c N_{c2}} s^2_{rc2}],$$

which reduces to,

$$Var(Y_D) = \sum_{c=1}^{k}[\overline{x}^2_{c1}\frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}}s^2_{rc1} + \overline{x}^2_{c2m}\frac{N_{c2}-m_c}{m_c N_{c2}}s^2_{rc2}]$$

Hence the proof. □

We observe that $MSE(Y_D)$ or equivalently, $Var(Y_D)$ vanishes as the stratum sample sizes in the first sampling and the second sampling phases increase.

7.4 Consistency

We define a sequence of point estimators of finite population total as $\{Y^*_D\}$. The sequence of point estimators $\{Y^*_D\}$ is said to be weakly consistent for Y_T if Y^*_D converges in probability to Y_T as the sample

Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) DOI: 10.7176/MTM
Vol.9, No.7, 2019

size becomes large (Cochran, 1977).

Theorem 7.3. *For a large population, and consequently a large sample size, the unbiased ratio-type estimator Y_D is a consistent estimator of the finite population total Y_T*

Proof. We shall use the Chebychev's inequality to prove the consistency of Y_D in estimating Y_T.

That is, we need to show that for every $\varepsilon > 0$,

$$\lim_{n_{c1}, m_c \to \infty} Pr\{|Y_D - Y_T| > \varepsilon\} = 0 \tag{28}$$

Now,

$$Pr\{|Y_D - Y_T| > \varepsilon\} \leq \frac{Var(Y_D^*)}{\varepsilon^2} = \frac{1}{\varepsilon^2} \sum_{c=1}^{k} [\overline{x}_{c1}^2 \frac{N_{c1} - n_{c1}}{n_{c1} N_{c1}} s_{rc1}^2 + \overline{x}_{c2m}^2 \frac{N_{c2} - m_c}{m_c N_{c2}} s_{rc2}^2] \tag{29}$$

Taking limits as $n_{c1} \to N_{c1}$, $m_c \to n_{c2}$ and that $n_{c2} \to N_{c2}$, the right hand side of Eq. (29) tends to zero.

Hence, $Y_D^* \xrightarrow{p} Y_T$, which is the condition for consistency. □

7.5 Confidence Interval for Population Total

Suppose in each stratum, the sample sizes for both phase I and phase II are large such that the sampled units tend in distribution to normal distribution, then the confidence interval of Y_T is given by

$$Y_D \pm Z_{\frac{\alpha}{2}} \sqrt{Var(Y_D)} \tag{30}$$

Where $Z_{\frac{\alpha}{2}}$ is the Z-normal variate to be chosen for given value of confidence co-efficient $(1 - \frac{\alpha}{2})$

8 Efficiency Comparison

In this section, we compare the MSE of Y_D and the estimator in literature developed under non-response and using SRSWOR scheme.

Theorem 8.1. *The estimator Y_D is more efficient than the usual ratio estimator constructed under non-response using stratified random sampling if the variance S_{Rcj}^2 is much smaller compared to S_{cdj}^2, where*

$S_{cdj}^2 = \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} (Y_{cij} - R_c X_{cij})^2$ *and* $R_c = \frac{\overline{Y}_c}{\overline{X}_c}$ *for* $c = 1, 2, ...k$ *and* $j = 1, 2$

Proof. We shall use the result of Rao (1986) about properties of the usual ratio estimator of population mean under non-response using SRSWOR.

Rao (1986) expressed the usual ratio estimator for population mean $\overline{Y}$ under non-response as

$$t_R = \frac{\overline{y}^*}{\overline{x}^*}\overline{X} = r^*\overline{X} \tag{31}$$

Where, $\overline{y}^* = w_1\overline{y}_1 + w_2\overline{y}_{2m}$, $\overline{x}^* = w_1\overline{x}_1 + w_2\overline{x}_{2m}$ and $r^* = \frac{\overline{y}^*}{\overline{x}^*}$

Rao (1986) expressed Bias and a large sample approximation to the MSE of t_R as

$$B_1 = \frac{(1-f)}{n\overline{X}}(RS_x^2 - S_{xy}) + W_2\frac{(h-1)}{n\overline{X}}(RS_{x2}^2 - S_{xy2}) \tag{32}$$

$$M_1 = \frac{(1-f)}{n}\sum_{j=1}^{2}\frac{(NW_j-1)}{(N-1)}S_{dj}^2 + W_2\frac{(h-1)}{n}S_{d2}^2 \tag{33}$$

Where, $S_d^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - RX_i)^2$ and $S_{dj}^2 = \frac{1}{N_j-1}\sum_{i=1}^{N_j}(Y_{ij} - RX_{ij})^2$ and $f = \frac{n}{N}$

From Eq. (33), we have a large sample approximation to the Mean Square Error of $\widehat{Y}_T$ as

$$M_1(\widehat{Y}_T) = N\frac{(1-f)}{f}\sum_{j=1}^{2}\frac{(N_j-1)}{(N-1)}S_{dj}^2 + N_2\frac{(h-1)}{f}S_{d2}^2 \tag{34}$$

We can expand Eq. (34) as

$$M_1(\widehat{Y}_T) = N\frac{(1-f)}{f}\frac{N_1-1}{N-1}S_{d1}^2 + [N\frac{(1-f)}{f}\frac{N_2-1}{N-1} + N_2\frac{h-1}{f}]S_{d2}^2 \tag{35}$$

But for large N, N_1 and N_2 such that $N-1 \approx N$, $N_1 - 1 \approx N_1$ and $N_2 - 1 \approx N_2$, then Eq. (35) reduces to

$$M_1(\widehat{Y}_T) = N\frac{(1-f)}{f}N_1S_{d1}^2 + [\frac{(1-f)}{f}N_2 + N_2\frac{(h-1)}{f}]S_{d2}^2 \tag{36}$$

Further, using the assumption that m is large and tends close to n_2 such that $h \to 1$, then Eq. (36) becomes,

$$M_1(\widehat{Y}_T) = \frac{(1-f)}{f}\{N_1S_{d1}^2 + N_2S_{d2}^2\} \tag{37}$$

Where S_{d1}^2 and S_{d2}^2 are the respective population variances in the responding and non-responding groups.

Under stratified random sampling, we can express Eq. (37) as

$$M_1(\widehat{Y}_T) = \sum_{c=1}^{k}\frac{N_c - n_c}{n_c}\{N_{c1}S_{c1}^2 + N_{c2}S_{c2}^2\} \tag{38}$$

Now, for Y_D, we have shown that $MSE(Y_D) = Var(Y_D)$ and using the proof under Corollary 1, we can express the population MSE af Y_D as

$$MSE(Y_D) = \sum_{c=1}^{k}[\overline{X}_{c1}^2\frac{N_{c1} - n_{c1}}{n_{c1}N_{c1}}S_{Rc1}^2 + \overline{X}_{c2}^2\frac{N_{c2} - m_c}{m_cN_{c2}}S_{Rc2}^2] \tag{39}$$

For Y_D to be more efficient than Y_T, we need to show that $MSE(Y_D) < M_1(\widehat{Y}_T)$. Or equivalently, we wish to find conditions under which efficiency of the proposed estimator is higher than that of the usual ratio estimator in literature under stratified random sampling. Since both are constructed using stratified random sampling, we shall only consider a particular stratum c, say.

From Eq. (38) and Eq. (39), we shall compare $\frac{N_{c1}}{n_c}(N_c - n_c)S_{cd1}^2$ and $\overline{X}_{c1}^2 \frac{N_{c1}-n_{c1}}{n_{c1}N_{c1}} S_{Rc1}^2$ for the responding group and $\frac{N_{c2}}{n_c}(N_c - n_c)S_{cd2}^2$ and $\overline{X}_{c2}^2 \frac{N_{c2}-m_c}{m_c N_{c2}} S_{Rc2}^2$ for the non-responding group.

But for a large m_c such that $m_c \approx n_{c2}$ then $\frac{N_{c2}-m_c}{m_c N_{c2}} \approx \frac{N_{c2}-n_{c2}}{n_{c2}N_{c2}}$

Therefore, in general, we compare $\frac{N_{cj}}{n_c}(N_c - n_c)S_{cj}^2$ and $\overline{X}_{cj}^2 \frac{N_{cj}-n_{cj}}{n_{cj}N_{cj}} S_{Rcj}^2$.

That is, we wish to show that

$$\frac{N_{cj}}{n_c}(N_c - n_c)S_{cj}^2 > \overline{X}_{cj}^2 \frac{N_{cj} - n_{cj}}{n_{cj}N_{cj}} S_{Rcj}^2 \tag{40}$$

Clearly, from Eq. (40), the inequality holds if the stratum variance of the ratios for the j^{th} group, S_{Rcj}^2, is much small compared to the stratum variance of the response variable for the j^{th} group, S_{cj}^2.

Hence the proof.

□

9 Conclusion

We have adopted the Hansen-Hurwitz sub-sampling method to construct a ratio-type estimator under non-response in the study variable in stratified random sampling scheme. From the asymptotic properties, we have observed that the proposed estimator is unbiased and consistent. From the efficiency comparison, we observed that the proposed estimator is more efficient than the usual ratio estimator constructed under non-response using stratified random sampling if the stratum variance of the ratios for the j^{th} group, S_{Rcj}^2, is much small compared to the stratum variance of the response variable for the j^{th} group, S_{cj}^2. We recommend that further study on the proposed estimator should be done to investigate whether it is a best linear unbiased estimator (BLUE) among a class of unbiased estimators for Y_T. Also, an empirical and/or simulation study should be done on the proposed estimator to verify its aforementioned properties.

References

[1] Chaudhary M. K., Malik S., Singh J. & Singh R. (2013). A General Family of Estimators for Estimating Population Mean in Systematic Sampling Using Auxiliary Information in the Presence of Missing Observations. *Jour. Raj. Stat. Assoc.*, 1(2), 1-8.

[2] Cochran W. G. (1977). *Sampling Techniques*. 3rd Edition. New York, John Wiley.

[3] Daroga, S. & Chaudhary F. (2002). *Theory and Analysis of Sample Survey Designs*. New Delhi, New Age International (P) Limited Publishers.

[4] Hansen M. H. & Hurwitz W. N. (1946). The Problem of Non Response in Sample Surveys. *Journal of the American Statistical Association*, 41, 517-529.

[5] Khoshnevisan M., Singh R., Chauhan P., Sawan N. & Smarandache F. (2007). A General Family of Estimators for Estimating Population Mean Using Known Value of Some Population Parameters. *Far East journal of Theoretical Statistics*, 22, 181-191.

[6] Kumar, S. (2012). Utilization of Some Known Population Parameters for Estimating Population Mean in Presence of Non-Response, *Pak.j.stat.oper.res.*, 8(2), 233-244.

[7] Oyoo D. & Ouma C. (2014). Estimation of Population Total in the Presence of Missing Values using a Modified Murthy's Estimator and the weight Adjustment Technique. *American Journal of Applied Mathematics and Statistics*. 2(3), 163-167.

[8] Rao, P. S. R. S. (1986). Ratio Estimation with sub sampling the non-respondents. *Survey Methodology*, 12(2), 217-230.

[9] Reddy, V. N. (1973). On ratio and product method of estimation. *Sankhya*, B, 35, 307-317.

[10] Shabbir, J. & Saghir, A. (2012). Estimation of Finite Population Mean in Stratified Random Sampling Using Auxiliary Attribute(s) under Non-Response. *Pak.j.stat.oper.res*, 8(1), 73-82.

[11] Singh, R. & Malik, S. (2014). Improved estimation of population variance using information on auxiliary attribute in simple random sampling. *Applied Mathematics and Computation*, 235, 43-49.

[12] Srivastava, S. K. (1967). An estimator using auxiliary information in sample surveys. *Calcutta Statistical Association Bulletin*, 16, 121-132.

[13] Walsh, J. E. (1970). Generalization of ratio estimates for population total. *Sankhya*, A, 32, 99-106.