Estimation of Finite Population Variance Using Two-Phase Sampling under Random Non-Response
Manoj K. Chaudhary¹, Gautam K. Vishwakarma and Anil Prajapati
Department of Statistics, Banaras Hindu University, Varanasi, India.

ABSTRACT
The present paper deals with the problem of estimating the finite population variance using two-phase sampling scheme in the presence of random non-response. In this paper, we have suggested some families of factor-type estimators of population variance utilizing the information on an auxiliary variable with unknown population variance. The properties of the suggested families of estimators have been discussed in detail. The optimum estimators of the suggested families have also been pioneered out. The theoretical results have been demonstrated through some real data sets. A simulation study has also been carried out to support the theoretical results.

© 2019 Elixir All rights reserved.

1. Introduction
Variation is an inherent phenomenon in our day to day life. There are several researchers who have made significant contributions in estimating the finite population variance. Das and Tripathi (1978) have suggested the variance estimation technique utilizing the information on an auxiliary variable with known population variance and coefficient of variation. Isaki (1983), Das (1988), Kadilar and Cingi (2006), Singh and Chandra (2008), Dubey and Sharma (2008), Gupta and Shabbir (2008), Singh and Solanki (2013) and others have contributed a lot in estimating the finite population variance.

The problem of non-response is a most burning issue in mail surveys. Sometimes, researchers found that the non-response has probabilistic nature in the mail surveys. Tracy and Osahan (1994) studied the effect of random non-response on the usual ratio estimator of the population mean under two different situations. Singh and Joarder (1998) presented the study of several estimators of finite population variance under two different situations of random non-response. Singh et al. (2012) have proposed some general classes of estimators of population variance under random non-response in survey sampling. Bandyopadhyay and Singh (2015) have suggested some estimators for estimating the ratio of population variances under two different realistic situations of complete response and random non-response.

The above recent works of estimation of population mean or population variance in the presence of random non-response have been studied under the assumption that either population mean or both population mean and population variance of the auxiliary variable are known. But, if the situation occurs where these parametric values of auxiliary variable are not known, it would be much difficult to estimate the parameters of study variable by utilizing the auxiliary information. Under such situations, one can adopt two-phase or double sampling scheme to estimate the parameters of study variable. In two-phase or double sampling scheme, it is advisable to draw a large preliminary sample for measuring the auxiliary character and then to draw a smaller sub-sample from it for collecting the information on the study variable. Double sampling scheme comes to be a powerful and cost effective technique for obtaining the reliable estimate in first-phase (preliminary) sample for the unknown population parameters of the auxiliary variable.

In this paper, we have proposed some families of estimators of finite population variance utilizing the information on an auxiliary variable with unknown population mean or/and population variance in the presence of random non-response. In order to propose the families, we have adopted two-phase sampling scheme. The expressions for the biases and mean square errors of the proposed families have been derived up to the first order of approximation. An empirical study has been carried out by considering some real data sets and a simulation viewpoint.

2. Sampling Strategy under Random Non-Response
Consider a population $P(U_1, U_2, ..., U_N)$ comprises of $N$ units. Let $X_0$ and $X_1$ be the study and auxiliary variables with respective population means $\bar{X}_0$ and $\bar{X}_1$. Let $x_{0i}$ and $x_{1i}$ be the observations on the $i^{th}$ unit in the population for the variables $X_0$ and $X_1$ respectively ($i = 1, 2, ..., N$). The objective of present research is to estimate the population variance of study variable $X_0$, i.e., $S_0^2 = (N - 1)^{-1} \sum_{i=1}^{N} (x_{0i} - \bar{X}_0)^2$ utilizing the information on the auxiliary...
variable \(X_i\) with unknown population variance \(S_i^2 = (N - 1)^{-1} \sum_{i=1}^{n} (x_{ni} - \bar{x}_i)^2\). In order to estimate \(S_0^2\), first a larger sample of size \(n\) is selected from the entire population of \(N\) units by the method of simple random sampling without replacement (SRSWOR) scheme and information is observed on auxiliary variable for estimating the parameter \(S_1^2\). Secondly, a smaller sub-sample of size \(n\) is selected from the first phase sample of \(n\) units by SRSWOR scheme and information is observed on both study and auxiliary variables. Let \(r = \frac{r}{r} + 1, \ldots, (n - 2)\) be the number of units in the second phase sample of size \(n\), on which information could not be collected for the study variable \(X_0\) due to random non-response. Thus, the observations for the variable on which random non-response occurs can be collected from the remaining \((n - r)\) units of the second phase sample. It is assumed that \(r\) should be less than \((n - 1)\), i.e. \(0 \leq r \leq (n - 2)\). Let \(p\) be the probability of a non-response among \((n - 2)\) cases of non-response. Then the discrete probability distribution of \(r\) is given by

\[
P(r) = \left(\frac{n - r}{nq + 2p}\right)^{n - r} \left(\frac{r}{r}ight)^{n - 2 - r} q^{n - 2 - r}
\]

where \(q = (1 - p)\). \(r = 0,1,2,\ldots, (n - 2)\) and \((n - 2)\) represents the total number of ways of obtaining \(r\) non-responses out of total \((n - 2)\) possible non-responses.

3. Suggested Families of Estimators

Our aim is to estimate the population variance \(S_0\) utilizing the information on an auxiliary variable with unknown population variance in the presence of random non-response. Thus, motivated by Singh and Shukla (1987), we now propose two different families of estimators of population variance \(S_0^2\) under the probability model given in equation (1) as

\[
T_{a} = \frac{1}{n - r} \sum_{i=1}^{n - r} \left( \frac{(A + C)x_{ni}^2 + fBx_{ni}^2}{(A + fB)x_{ni}^2 + Cx_{ni}^2} - \bar{x}_0^2 \right)
\]

and

\[
T_{a} = \frac{1}{n - r} \sum_{i=1}^{n - r} \left( \frac{(A + C)x_{ni}^2 + fBx_{ni}^2}{(A + fB)x_{ni}^2 + Cx_{ni}^2} - \bar{x}_0^2 \right)
\]

where

\[
s_i^2 = \frac{1}{n - r} \sum_{i=1}^{n - r} \left( x_{ni} - \bar{x}_i \right)^2 \quad \bar{x}_i = \frac{1}{n - r} \sum_{i=1}^{n - r} x_{ni} \quad \bar{x}_0 = \frac{1}{n - r} \sum_{i=1}^{n - r} x_{oi} \quad i = 1, 2, \ldots, \alpha
\]

\[
A = (\alpha - 1)(\alpha - 2) \quad B = (\alpha - 1)(\alpha - 3) \quad C = (\alpha - 2)(\alpha - 3)(\alpha - 4)
\]

for \(\alpha > 0\) and \(r = \frac{n}{N}\).

Remark: The families proposed in equations (2) and (3) can generate a number of well known existing and some other estimators of population variance \(S_0^2\) under random non-response for the suitable choices of \(\alpha\). For instance, if \(\alpha = 1\), \(\alpha = 2\), \(\alpha = 3\) and \(\alpha = 4\), we respectively get ratio-type estimator, product-type estimator, dual to ratio-type estimator and usual variance estimator under random non-response.

4. Properties of the Proposed Families of Estimators

To obtain the biases and mean square errors (MSE) of the proposed families \(T_{a}^r\) and \(T_{a}^r\), we use large sample approximation theory. Let us assume

\[
s_i^2 = \frac{1}{N} \left( 1 + e_i \right) \quad s_1^2 = \frac{1}{N} \left( 1 + e_1 \right) \quad s_i^2 = \frac{1}{N} \left( 1 + e_i \right) \quad s_i^2 = \frac{1}{N} \left( 1 + e_3 \right) \quad |e_i| < 1 \quad i = 0,1,2,3
\]

such that \(E(e_0) = E(e_1) = E(e_2) = E(e_3) = 0\).

\[
E(e_0^2) = \left( \frac{1}{nq + 2p} - \frac{1}{N} \right) \left( \lambda_{a4} - 1 \right) \quad E(e_1^2) = \left( \frac{1}{nq + 2p} - \frac{1}{N} \right) \left( \lambda_{a4} - 1 \right)
\]

\[
E(e_2^2) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( \lambda_{a4} - 1 \right) \quad E(e_3^2) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( \lambda_{a4} - 1 \right) \quad E(e_0 e_1) = \left( \frac{1}{nq + 2p} - \frac{1}{N} \right) \left( \lambda_{a4} - 1 \right)
\]

\[
E(e_0 e_2) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( \lambda_{a2} - 1 \right) \quad E(e_1 e_2) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( \lambda_{a4} - 1 \right)
\]

\[
E(e_0 e_3) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( \lambda_{a2} - 1 \right) \quad E(e_1 e_3) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( \lambda_{a4} - 1 \right)
\]

\[
E(e_2 e_3) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( \lambda_{a2} - 1 \right) \quad E(e_0 e_3) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( \lambda_{a4} - 1 \right)
\]
where \( \lambda_{ts} = \frac{\mu_{ts}}{\mu_{22}^{1/2}} \) and \( \mu_{ts} = (N - 1)^{-1} \sum_{i=1}^{N} (x_{0i} - \bar{X}_0) (x_{2i} - \bar{X}_1) \).

4.1 Properties of the Family \( T_{\alpha}^{**} \)

Expressing equation (2) in terms of \( e_0, e_1 \) and \( e_2 \), we get

\[
T_{\alpha}^{**} = S_0^2 \left[ 1 + e_0 \right] \left[ \frac{1 + \phi_1(\alpha)}{1 + \phi_2(\alpha)} \left( \frac{e_1 - e_2}{1 + e_2} \right) \right] \]

where \( \phi_1(\alpha) = \frac{f_B}{A + f_B + C} \) and \( \phi_2(\alpha) = \frac{C}{A + f_B + C} \).

Expanding equation (4) and neglecting the terms having powers of \( e_0, e_1 \) and \( e_2 \) greater than two, we get

\[
T_{\alpha}^{**} - S_0^2 = S_0^2 \left[ e_0 - \phi(\alpha) \left( e_1 - e_2 + e_2 + e_0 e_1 - e_0 e_2 - e_1 e_2 \right) \right] \]

where \( \phi(\alpha) = \phi_3(\alpha) - \phi_2(\alpha) = \frac{C - f_B}{A + f_B + C} \).

Taking expectation on both the sides of equation (5), we get the bias of \( T_{\alpha}^{**} \) up to the first order of approximation as

\[
\text{Bias}(T_{\alpha}^{**}) = S_0^2 \left[ f' \phi(\alpha) \phi_3(\alpha) C_1^2 - f' \phi(\alpha) \rho_{01} C_0 C_1 \right] \]

where \( f' = f'' - f_2 = \left( \frac{1}{nq + 2p} - \frac{1}{n} \right) \) and \( f'' = \left( \frac{1}{nq + 2p} - \frac{1}{N} \right) \)

\[
C_1 = \sqrt{\lambda_{04} - 1} \text{ and } \rho_{01} = \frac{1}{\sqrt{(\lambda_{04} - 1)(\lambda_{04} - 1)}} \]

Squaring both the sides of equation (5) and then taking expectation on ignoring the terms having powers of \( e_0, e_1 \) and \( e_2 \) higher than two, we get the MSE of \( T_{\alpha}^{**} \) up to the first order of approximation as

\[
\text{MSE}(T_{\alpha}^{**}) = S_0^4 \left[ f' \phi^2(\alpha) C_1^2 - 2 f' \phi(\alpha) \rho_{01} C_0 C_1 \right] \]

To get the optimum value of \( \alpha \) so that the MSE of \( T_{\alpha}^{**} \) would attain its minimum, we differentiate equation (7) with respect to \( \alpha \) and equate the derivative to zero. Thus, the normal equation reduces to

\[
\phi(\alpha) = \frac{\rho_{01} C_0}{C_1} \]

The above equation provides three real roots of \( \alpha \) for which one can get the minimum MSE of \( T_{\alpha}^{**} \).

Putting the value of \( \phi(\alpha) \) from equation (8) into equation (7), we get the expression for minimum MSE of \( T_{\alpha}^{**} \) as

\[
\text{MSE}(T_{\alpha}^{**})_{\text{min}} = S_0^4 \left[ f' - f \rho_{01} C_0^2 \right] \]

4.2 Properties of the Family \( T_{\alpha} \)

Expressing equation (3) in terms of \( e_0, e_2, e_3 \) and neglecting the terms having powers of \( e_0, e_2, e_3 \) greater than two, we get

\[
T_{\alpha}^{**} - S_0^2 = S_0^2 \left[ e_0 - \phi(\alpha) \left( e_2 + e_3 + e_0 e_2 + e_0 e_3 - e_2 e_3 \right) \right] \]

Now, taking expectation on both the sides of equation (10), we get the bias of \( T_{\alpha}^{**} \) up to the first order of approximation as

\[
\text{Bias}(T_{\alpha}^{**}) = S_0^2 \left[ f_3 \phi(\alpha) \phi_3(\alpha) C_1^2 - f_3 \phi(\alpha) \rho_{01} C_0 C_1 \right] \]

where \( f_3 = f_1 - f_2 = \left( \frac{1}{n} - \frac{1}{n} \right) \) and \( f_1 = \left( \frac{1}{n} - \frac{1}{N} \right) \).

To get the MSE of \( T_{\alpha}^{**} \) up to the first order of approximation, we square both the sides of equation (10) and neglect the terms having powers of \( e_0, e_2 \) and \( e_3 \) higher than two. Thus, the required expression for MSE is given as

\[
\text{MSE}(T_{\alpha}^{**}) = S_0^4 \left[ f' \phi^2(\alpha) C_1^2 - 2 f_3 \phi(\alpha) \rho_{01} C_0 C_1 \right] \]

In order to obtain the minimum MSE of \( T_{\alpha}^{**} \), we differentiate the equation (12) with respect to \( \alpha \) and equate the derivative to zero. Thus, we have

\[
\phi(\alpha) = \frac{\rho_{01} C_0}{C_1} \]
which is the same as equation (8) and provides three real roots for $\alpha$. Thus, the expression for minimum MSE of $T^{**}_\alpha$ is given as

$$MSE(T^{**}_\alpha)_{min} = S_0^2[f^* - f_3\rho_0^2)^2]$$

(13)

5. Empirical Study

To realize the facts obtained in this paper and to examine the behaviour of the families of estimators, it is essential to illustrate whatever has been discussed in previous sections with some numerical data. To support the theoretical results, we have done the empirical study through some real data sets as well as a simulation viewpoint.

5.1 Real Data Sets

Data Set-1:

We have used the data considered by Sukhatme (1970). The details are given below:

$$N = 34 \cdot n = 20 \cdot n = 12$$

$$\lambda_22 = 2.433161 \cdot \lambda_{04} = 3.4426 \cdot C_0 = 1.56289040 \cdot C_1 = 1.47819885$$

The Table 1 shows MSE and percentage relative efficiency (PRE) of the families $T^{***}_\alpha$ and $T^{*}_\alpha$ for the different choices of non-response probability $\rho(= 0.05, 0.10, 0.15, 0.20)$. PRE is computed with respect to usual variance estimator under random non-response.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\alpha$</th>
<th>$T^{***}_\alpha$</th>
<th>$T^{*}_\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MSE</td>
<td>PRE</td>
</tr>
<tr>
<td>0.05</td>
<td>1</td>
<td>61862640.47</td>
<td>121.82</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>17545011.11</td>
<td>42.95</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>82584945.25</td>
<td>91.25</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>75361054.7</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{opt}$</td>
<td>56735899.06</td>
<td>132.83</td>
</tr>
<tr>
<td>0.10</td>
<td>1</td>
<td>65595274.66</td>
<td>122.78</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>19131137.9</td>
<td>42.09</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>85533868.74</td>
<td>90.97</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>80537369.56</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{opt}$</td>
<td>59920218.7</td>
<td>134.41</td>
</tr>
<tr>
<td>0.15</td>
<td>1</td>
<td>69683397.8</td>
<td>123.71</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>208724643.3</td>
<td>41.30</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>95049356.33</td>
<td>90.70</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>86206666.8</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{opt}$</td>
<td>63407806.85</td>
<td>135.96</td>
</tr>
<tr>
<td>0.20</td>
<td>1</td>
<td>71480333.28</td>
<td>124.62</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>227857499.5</td>
<td>40.57</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>102216392.7</td>
<td>90.44</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>92442893.74</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{opt}$</td>
<td>67244153.77</td>
<td>137.47</td>
</tr>
</tbody>
</table>

Data Set-2:

Now, we have used another data set considered by Shukla and Thakur (2008). In this data set, a population of 200 units was considered. The parameters and other details of the population are given below:

$$N = 200 \cdot n = 20 \cdot n = 12 \cdot \lambda_22 = 42.4851 \cdot \lambda_{04} = 3.74 \cdot \lambda_{90} = 2.56 \cdot C_0 = 1.248999600, C_1 = 1.655294536$$

The Table 2 also shows MSE and PRE of the families $T^{***}_\alpha$ and $T^{*}_\alpha$ for the different choices of non-response probability $\rho(= 0.05, 0.10, 0.15, 0.20)$. PRE is computed as same as in Table 1.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\alpha$</th>
<th>$T^{***}_\alpha$</th>
<th>$T^{*}_\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MSE</td>
<td>PRE</td>
</tr>
<tr>
<td>0.05</td>
<td>1</td>
<td>3839.63</td>
<td>131.94</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>14317.49</td>
<td>35.38</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4533.55</td>
<td>111.75</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5066.12</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{opt}$</td>
<td>3356.04</td>
<td>150.96</td>
</tr>
<tr>
<td>0.10</td>
<td>1</td>
<td>3952.79</td>
<td>134.35</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>15551.26</td>
<td>34.15</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4720.92</td>
<td>112.49</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5310.44</td>
<td>100.00</td>
</tr>
</tbody>
</table>
There is an increase in non-response probability. Therefore, we have generated 2000 values for each of the variables. The procedure of selecting the second phase sample of size \( S \), dropping the non-response probability, we have selected a first phase sample of \( n \), and \( S \) and \( \rho \).

From Table 1 and Table 2, it is revealed that the optimum estimators of the proposed families perform best among all other estimators. It is also revealed that the MSE of the estimators increases with increase in non-response probability.

### 5.2 Simulation Viewpoint

In order to demonstrate the theoretical results through a simulation analysis, we define the study variable \( X_0 \) using the transformation \( X_0 = \rho X_1 + \sqrt{1 - \rho^2} X_2 \), where \( X_1 \) and \( X_2 \) are the random variables from a normal population i.e. \( X_1 \sim \mathcal{N}(5,2) \) and \( X_2 \sim \mathcal{N}(5,2) \). Therefore, we have generated 2000 values for each of the variables \( X_1 \) and \( X_2 \) using R software and then realized the 2000 values for the study variable \( X_0 \) using the transformation. Further, \( X_1 \) has been considered as the auxiliary variable. Now, we have selected a first phase sample of \( n = 500 \) units from the population of \( N = 2000 \) units using SRSWOR scheme and then selected a second phase sample of \( n = 100(5)200 \) units from the first phase sample using SRSWOR scheme. In the next step, 20\% units of the second phase sample were randomly dropped down. The dropped down units have been treated as non-responding units. At this stage, we have computed the estimates of \( S_0^2 \) using the families of estimators \( T_{\alpha}^{**} \) and \( T_{\alpha}^{*} \). The procedure of selecting the second phase sample of size \( n \), dropping down the non-responding units and computing the estimates of \( S_0^2 \) was repeated 500 times. Finally, we computed the average MSE (AM) of the families \( T_{\alpha}^{**} \) and \( T_{\alpha}^{*} \) using the following formulae:

\[
AM(T_{\alpha}^{**}) = \frac{1}{500} \sum_{k=1}^{500} (T_{\alpha k}^{**} - S_0^2)^2
\]

\[
AM(T_{\alpha}^{*}) = \frac{1}{500} \sum_{k=1}^{500} (T_{\alpha k}^{*} - S_0^2)^2 ; \quad k = 1,2,...,500
\]

Table 3 represents the average MSE of the families \( T_{\alpha}^{**} \) and \( T_{\alpha}^{*} \) for \( \rho = 0.80 \) and \( n = 100(5)200 \) at \( p = 0.20 \).

### Table 3. Average MSE of the Families \( T_{\alpha}^{**} \) and \( T_{\alpha}^{*} \).

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( T_{\alpha}^{**} )</th>
<th>( T_{\alpha}^{*} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>100</td>
<td>150</td>
</tr>
<tr>
<td>1</td>
<td>0.01562</td>
<td>0.00649</td>
</tr>
<tr>
<td>2</td>
<td>0.17346</td>
<td>0.10085</td>
</tr>
<tr>
<td>3</td>
<td>0.02688</td>
<td>0.01494</td>
</tr>
<tr>
<td>4</td>
<td>0.04677</td>
<td>0.02766</td>
</tr>
<tr>
<td>( \alpha_{opt} )</td>
<td>0.01209</td>
<td>0.00462</td>
</tr>
</tbody>
</table>

### 6. Conclusions

We have proposed some families of factor-type estimators for estimating the finite population variance under random non-response. In order to propose the families of estimators, we have utilized the information on an auxiliary variable with unknown variance. The proposed families can produce random non-response version of a number of well known existing and some other estimators of population variance. The optimum estimators of the proposed families have also been pioneered out. The properties of the families of estimators have been discussed in detail. To support the theoretical results, an empirical study has also been carried out through some real data sets and a simulation point of view. The Table 1 and Table 2 show that the optimum estimators of the proposed families provide best estimates among all other existing estimators. The similar results have been found in Table 3. From Table 1 and Table 2, it is also observed that the MSE of the estimators increases with increase in random non-response probability. Such a result is intuitively expected.

### References

15. Tracy, D. S., Osahan, S. S.: Random non-response on study variable versus on study as well as auxiliary variables, Statistica, 54, 163-168 (1994)