Int. J. Agricult. Stat. Sci. Vol. 17, No. 2, pp. 733-739, 2021
DocID: https://connectjournals.com/03899.2021.17.733
www.connectjournals.com/ijass
ISSN : 0973-1903, e-ISSN : 0976-3392
ORIGINAL ARTICLE

PREDICTION APPROACH ESTIMATOR USING AUXILIARY INFORMATION UNDER TWO-PHASE SAMPLING

Nitin Varshney, Tauqueer Ahmad*, Anil Rai, Ankur Biswas and Prachi Misra Sahoo
ICAR-Indian Agricultural Statistics Research Institute, New Delhi - 110 012, India.
E-mail: tauqueer.khan01@gmail.com

Abstract: Sample survey is a cost effective mean to collect reliable information about a finite population. There are various sampling methodologies, among them two-phase sampling is generally used for estimating population mean or total under the two different situations. First, when the information of the auxiliary variable is not readily available and the other condition is when it is vey expensive to gather information on characteristic under study y, but it is comparatively cheaper to gather information on the variables which are highly correlated with the characteristic under study. In large scale surveys, two-phase sampling approach is proposed in order to reduce the number of sampled units which require the more expensive objective methods. Prediction approach is applied to predict the non-sampled units in surveys. In the large preliminary sample (first phase sample) of two-phase sampling, there are total n'- n non-sampled units having auxiliary information, so there is a need to develop an estimator based on prediction approach under finite population. In the present study, we have proposed a new estimator of finite population total based on prediction approach in the context of two-phase sampling.

***Key words*:** Two-phase sampling, Auxiliary information, Prediction approach.

Cite this article

Nitin Varshney, Tauqueer Ahmad, Anil Rai, Ankur Biswas and Prachi Misra Sahoo (2021). Prediction Approach Estimator Using Auxiliary Information Under Two-Phase Sampling. *International Journal of Agricultural and Statistical Sciences.* DocID: https://connectjournals.com/03899.2021.17.733

1. Introduction

Sampling which is a technique of selection of a part of an aggregate to represent the whole is frequently used in everyday life in all kinds of investigations. Sample survey is a cost-effective technique to collect information about a finite population with one of the main aims to develop estimates of means and totals of several characteristics associated with a finite population [Cochran (1977)]. In many situations, estimates are necessary for the population as well as it is also required for various sub-populations. In survey literature, many estimation procedures require some advanced knowledge of auxiliary information. For instance, ratio, regression or product estimators require the availability of auxiliary information [Hidiroglou and Sarndal (1998)]. Kadilar and Cingi (2004) discussed ratio estimators in simple random sampling. Subzar *et al.*(2017) enhanced the ratio type estimators for estimating population mean using auxiliary information in survey sampling. Sarndal *et al.* (1992) discussed model assisted survey sampling in detail. Use of auxiliary information often increases the efficiency of such estimators [Ahuja and Misra (2020)]. When the auxiliary variable's information is not available, we first select a large preliminary sample to study the auxiliary variable whereas a small sub-sample is selected to observe the character under study and auxiliary variable. This sampling method is popularly known as double sampling or two-phase sampling [Hidiroglou (2001)]. Choudhury and Singh (2018) developed a general class of estimators for finite population mean using auxiliary variable in double sampling.

Neyman (1938) developed the theory of two-phase sampling, as an alternative to simple random sampling. It can be a cost efficient technique in large-scale surveys. It is common in surveys to use two-phase sampling when it is relatively inexpensive to draw a large 1st phase sample for which auxiliary variable

Author for correspondence* ***Received* *January 10, 2021* ***Revised*** *May 26, 2021* ***Accepted*** *September 27, 2021*

correlated with the characteristic of interest, alone is observed. A 2nd phase subsample of the initial 1st phase sample is then drawn and both auxiliary as well as characteristic of interest are measured. For example, If we do a survey to estimate the wheat yield of a given area of India, we can take a large sample of n' farms (1st phase sample) to estimate the total area (auxiliary information) under wheat cultivation and a sub-sample of *n* farms (2nd phase sample) to determine the actual wheat yield. Misra (2018) developed improved double sampling estimator of population mean using auxiliary information.

2. Prediction Approach in Sampling y_i

Let us consider that the number of units in the finite population is known. It is also known that with each unit, a number y_i is associated. As per the basic sampling theory, the main difficulty is to select few units of the population as a sample to observe the $y_i's$ for the sampling units, and then we will use those samples to estimate the value of some function $h(y_1, y_2, ..., y_N)$ of all the $y_i's$ in the population. The function $h(y_1, y_2, ..., y_N)$ can be a simple combination of $y_i's$ s like their total or mean or may be something more complex like quantile. In prediction approach population observations $y_1, y_2, ..., y_N$ are treated as realized values of random variables $Y_1, Y_2, ..., Y_N$. After the sample has been observed, it is necessary to predict a function of the non-sampled *Y's* by estimating $h(y_1, y_2, ..., y_N)$. So, prediction approach [Valliant *et al.* (2000)] is used to predict those non-sampled units of the population [Royal (1970)]. Royal (1978) discussed prediction theory in finite population sampling. Panda (2015) proposed an efficient predictive approach to estimation in two-phase sampling. Bandyopadhyay and Singh (2016) proposed prediction based estimation of population mean in the context of two-phase sampling. Misra *et al*. (2017) used information of auxiliary variable for estimating population mean and they developed an improved ratio-type predictive estimator. Let, the population total be $T = \sum_s y_i + \sum_r y_i$, where the first sum is over sampled part *s* consisting of *n* sampling units which is known and the second one is over non-sampled part *r* which must be predicted based on the observed sample. So we can say that estimation of population total means predicting the value of non-sampled units *i.e.* $\sum_r y_i$. Then, by the usual predictive approach, an estimator $\hat{T}$ of *T* can be written as

$$\hat{T} = \sum_s y_i + \sum_r \hat{y}_i \tag{1}$$

3. Proposed Estimator under Two-Phase Sampling

Prediction approach is applied to develop estimators of population parameters using prediction of the non-sampled units in surveys based on sampled units. In two-phase sampling, there are $n' - n$ non-sampled units having auxiliary information, whereas *n* sampled units having information on auxiliary as well as study variable. Thus, under finite population framework, Prediction approach based estimator can be developed under two-phase sampling by predicting the $n' - n$ non-sampled units based on *n* sampled units. In this present study, a generalized estimator for estimation of finite population total has been proposed under two-phase sampling by using design based as well as prediction based approaches.

Let us consider a finite population of size *N* having observations on study variable as $y_1, y_2, ..., y_N$. In the 1st phase of the two-phase sampling design, a preliminary large sample $s^{(1)}$ of size n' is drawn by a probability sampling design *i.e.* SRSWOR, where $\pi_i' = P\left(i \in s^{(1)}\right)$ is the known 1st phase inclusion probability of *i*th sampling unit, which reduces to $\pi_i' = n'/N$ under SRSWOR. In the 2nd phase of sampling, a smaller sample $s^{(2)}$ of size *n* is selected from the 1st phase sample $s^{(1)}$ by SRSWOR, where $\pi_{i|s^{(1)}} = P(i \in s^{(2)} \mid s^{(1)}) = n/n'$ is the 2nd phase conditional inclusion probability of *i*th sampling unit, given $s^{(1)}$. Let, the character under study is population total (*Y*) *i.e.* $Y = \sum_{i=1}^{N} Y_i$.

A model, which expressed the relationships between

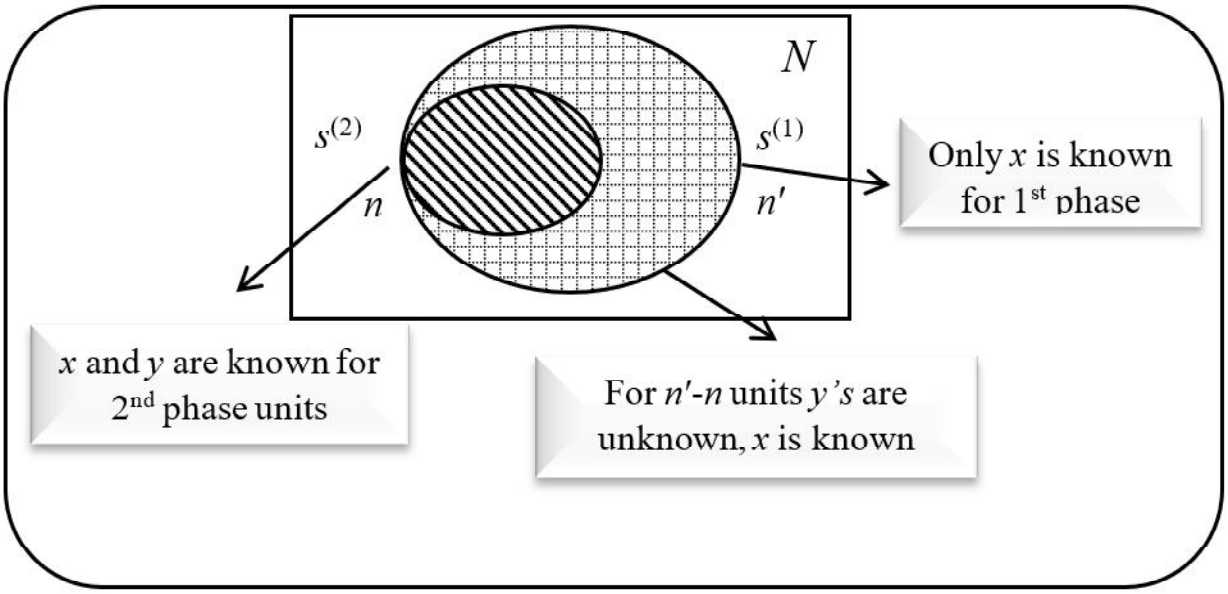

Fig. 1: Pictorial representation of Two-phase sampling

the random variables, is used for their joint probability distributions. The predictions are made on the basis of the these type of models. Regression model is used to predict y values of unobserved $n'-n$ units. Regression model is fitted on the basis of available x and y at 2nd phase to predict y values of unobserved $n'-n$ units.

$$\mathbf{y}_{s_2} = \mathbf{x}_{s_2}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where, $\varepsilon \sim N\left(0, \sigma^2 \mathrm{I}\right)$ and other terms are as following,

$$\mathrm{y}_{s_2} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathrm{x}_{s_2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{ and } \beta = \begin{bmatrix} 0 \\ \beta \end{bmatrix}$$

$$\mathrm{x}_{s_2}\beta = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} 0 \\ \beta \end{bmatrix} = \begin{bmatrix} \beta x_1 \\ \beta x_2 \\ \vdots \\ \beta x_n \end{bmatrix} = \beta \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Now the error term will become

$$\varepsilon = \mathrm{y}_{s_2} - \mathrm{x}_{s_2}\beta \tag{3}$$

According to the principle of least squares, we have to estimate β so that the sum of squares due to errors is minimum, *i.e.*

$$\hat{\beta} = \left(\mathrm{x}_{s_2}^{'}\mathrm{x}_{s_2}\right)^{-1}\mathrm{x}_{s_2}^{'}\mathrm{y}_{s_2} \tag{4}$$

Model based prediction approach entails the problem of estimating the population total is equivalent to predicting the value of the sum of non-sampled units. So the predicted values of the non-sampled units will be as follows.

$$\hat{\mathrm{y}}_{s_1-s_2} = \mathrm{x}_{s_1-s_2}\hat{\beta} \tag{5}$$

Now on putting the value of $\hat{\beta}$ from the Equation (4), we will get the predicted values of the sum of non-sampled units as

$$\hat{\mathrm{y}}_{s_1-s_2} = \mathrm{x}_{s_1-s_2}\left(\mathrm{x}_{s_2}^{'}\mathrm{x}_{s_2}\right)^{-1}\mathrm{x}_{s_2}^{'}\mathrm{y}_{s_2} \tag{6}$$

As we know that our character under study at first phase (n' level) is the sum of all sampled as well as non-sampled units, which are being predicted by Equation (6), so our character under study will be

$$\mathrm{y}_{s_1} = \begin{bmatrix} \mathrm{y}_{s_2} \\ \hat{\mathrm{y}}_{s_1-s_2} \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{y}_{s_2} \\ \mathrm{x}_{s_1-s_2}\left(\mathrm{x}_{s_2}^{'}\mathrm{x}_{s_2}\right)^{-1}\mathrm{x}_{s_2}^{'}\mathrm{y}_{s_2} \end{bmatrix} \tag{7}$$

Table 1: Comparison of the proposed estimator with respect to other traditional estimators of population total on the basis of % Relative Bias.

Sample Sets (N-n')	HT estimator	Ratio estimator	Regression Estimator	Proposed Estimator
1000-100	–0.004	–0.008	–0.008	0.005
1000-200	–0.017	–0.020	–0.020	–0.010
1000-300	0.039	0.011	0.010	0.019
1000-400	0.004	–0.001	–0.001	0.007
1500-150	0.044	–0.002	–0.002	0.009
1500-300	–0.006	–0.002	–0.002	0.006
1500-450	0.015	–0.025	–0.025	–0.017
1500-600	0.009	0.001	0.001	0.008
2000-200	0.047	0.031	0.032	0.042
2000-400	0.009	–0.001	–0.001	0.009
2000-600	–0.016	–0.004	–0.004	0.005
2000-800	0.007	0.004	0.004	0.011

Table 2: Comparison of the proposed estimator with respect to other traditional estimators of population total on the basis of RRMSE.

Sample Sets (N-n')	HT estimator	Ratio estimator	Regression Estimator	Proposed Estimator
1000-100	1.130	0.551	0.551	0.548
1000-200	0.769	0.430	0.430	0.429
1000-300	0.613	0.394	0.394	0.393
1000-400	0.538	0.360	0.360	0.360
1500-150	0.888	0.445	0.445	0.443
1500-300	0.604	0.341	0.341	0.340
1500-450	0.508	0.306	0.306	0.305
1500-600	0.433	0.289	0.289	0.288
2000-200	0.807	0.376	0.376	0.375
2000-400	0.552	0.290	0.290	0.290
2000-600	0.419	0.246	0.246	0.245
2000-800	0.365	0.230	0.230	0.230

Under this study, a prediction approach based estimator has been proposed under Two-phase sampling design by predicting $n'-n$ non-sampled units based on n sampled units. First, a design based estimator of population total (Y) based on 1st phase sample $s^{(1)}$ is given by

$$\hat{T} = \sum_{i=1}^{n'} \frac{y_i}{\pi_i'} = \frac{N}{n'} \sum_{i=1}^{n'} y_i = \frac{N}{n'} y' \tag{8}$$

Sample total of 1st phase sample $y' = \sum_{i=1}^{n'} y_i$ consists of two parts *viz*. 2nd phase sample in $s^{(2)}$ and the remaining units in the set $s^{(1)} - s^{(2)}$ which are unobserved. Using prediction approach, an estimator of $y' = \sum_{i=1}^{n'} y_i$ based on 2nd phase sample $s^{(2)}$ and predicted values of the set $s^{(1)} - s^{(2)}$ as given in Equation (6) is given by

$$\hat{y}' = \sum_{s_2} y_i + \sum_{s_1 - s_2} \hat{y}_i$$

$$= \sum_{s_2} y_i + \hat{\beta} \sum_{s_1 - s_2} x_i$$

$$= n\bar{y}_n + \hat{\beta}\left(n'\bar{x}_{n'} - n\bar{x}_n\right)$$

where,

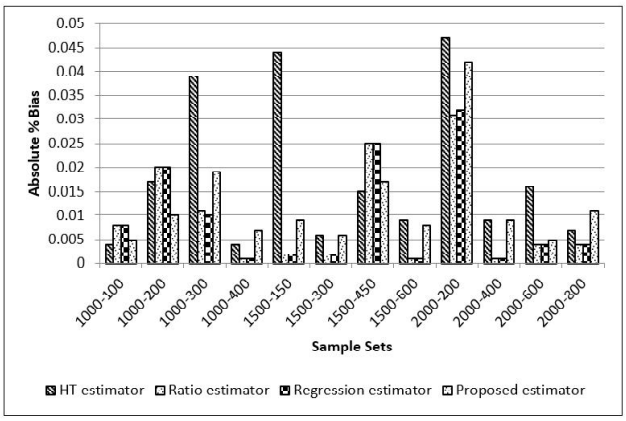

Fig. 2: Comparison of the estimators of population total on the basis of absolute % relative bias

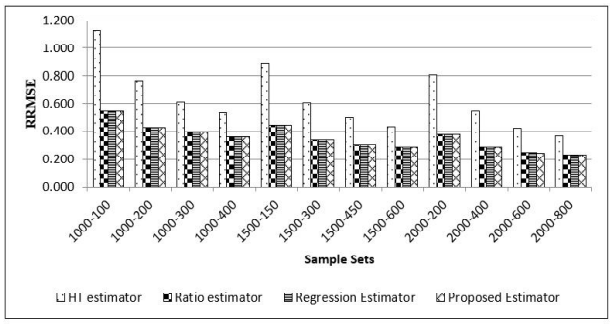

Fig. 3: Comparison of the estimators of population total on the basis of RRMSE

Table 3: Parameters of generated population under simulation study

Parameter	β	σ^2	ρ	N
Value	3	2	0.91	5000

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Therefore, using the above shown prediction estimator of y' in the Equation (8), the proposed estimator of the population total (Y) under two-phase sampling is given by

$$\hat{T}_{pred} = \frac{N}{n'}\hat{y}' = \frac{N}{n'}\left[n\bar{y}_n + \hat{\beta}\left(n'\bar{x}_{n'} - n\bar{x}_n\right)\right] \quad (9)$$

4. Bias of the Proposed Estimator

In Statistics, the difference between the estimator's expected value and the actual value of the parameter (which we wants to estimate) is known as the bias of an estimator. In this study, our parameter is population total, so bias of the proposed estimator will be calculated as

$$Bias\left(\hat{T}_{pred}\right) = E\left(\hat{T}_{pred}\right) - N\bar{Y}_N \quad (10)$$

After calculating the expected value of the proposed estimator

$$E\left(\hat{T}_{pred}\right) = E_1E_2E_M\left(\hat{T}_{pred}\right) = \frac{N}{n'}\{n\bar{Y}_N + \beta\left(n' - n\bar{X}_N\right) \quad (11)$$

Bias of the proposed estimator will be

$$Bias\left(\hat{T}_{pred}\right) = -\frac{\left(n'-n\right)}{n'}\left(\bar{Y}_N - \beta\bar{X}_N\right). \quad (12)$$

5. Existing Estimators Under Two-Phase Sampling

We have reviewed some existing estimators for estimating population total in case of Two-phase sampling in order to compare the performance of the proposed estimator.

5.1 Horvitz Thompson (H-T) estimator

When the sample is drawn using a probability sampling design, an unbiased estimator of $Y = \sum_{i=1}^{N} y_i$ given by

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \quad (13)$$

where, π_i is the 1st order inclusion probability. This estimator is known as H-T estimator, which is originally developed by Horvitz and Thompson (1952). Since under Two-phase sampling, there are Two-phases, thus, inclusion probability, π_i is considered as product of inclusion probabilities of both phases. This estimator is also called expansion estimator.

5.2 Ratio estimator

In Ratio estimator, to estimate the population mean of the character under study the knowledge of population mean of auxiliary variable is required. If the information on the auxiliary variable is not available, then, we use Two-phase sampling to estimate the unknown population mean of auxiliary variable on the basis of large preliminary sample of size n'. So, the ratio estimator in context of Two-phase sampling is defined as

$$\bar{y}_{rd} = \frac{\bar{y}_n}{\bar{x}_n}\bar{x}_{n'} \tag{14}$$

5.3 Regression estimator

It was found that ratio type estimators performs better if the line of regression (Y on X) is linear and passes through the origin. In many situations, it is found that sometimes, line of regression of Y on X is linear but it does not pass through the origin. Under such situations, it seems more appropriate to use regression type estimators. The regression estimator in context of Two-phase sampling is defined as

$$\bar{y}_{lr} = \bar{y}_n + b\left(\bar{x}_{n'} - \bar{x}_n\right) \tag{15}$$

where, b is the regression coefficient of y on x estimated from the 2nd phase sample.

6. Simulation Study

In order to study the statistical behaviour of the proposed estimator, a simulation study has been undertaken. For this study, first a bivariate population of size 5000 units is generated with the help of following model. Let y_i denotes the study variable, x_i denotes the auxiliary variable, β denotes the regression coefficient of y on x and ϵ_i denotes the error term. Then, the model can be written as

$$y_i = \beta x_i + \epsilon_i \tag{16}$$

where, ϵ_i are normally distributed with mean zero and constant variance σ_e^2, *i.e.* $\epsilon_i \sim N\left(0, \sigma_e^2\right)$ independent of x_i which also follows normal distribution $x_i \sim N(10,1)$. The obtained values of x_i and ϵ_i will thus be utilized to get the corresponding y_i values using the model. The parameters chosen for the population are given in Table 3:

Thus, for obtaining a two-phase random sample (SRSWOR at each phase), first, a 1st phase sample $s^{(1)} = \{(x_i); i = 1,2,...,n'\}$ has been generated and from this 1st phase sample $s^{(1)}$, a 2nd phase sample $s^{(2)} = \left\{\left(x_i, y_i\right); i = 1,2,...,n; i \in s^{(2)}\right\}$ has been generated. From the two phase sample, all the estimators discussed earlier are calculated. This process is repeated independently 1000 times.

To compare the statistical performance of the proposed estimator with the existing estimators in the context of two-phase sampling % relative bias (%RB) and relative root mean square (RRMSE) have been considered.

i. **% Relative Bias (RB):** % Relative Bias is calculated using following formula

$$\%\ Relative\ Bias = \frac{1}{S}\sum_{j=1}^{S}\left[\frac{\hat{T}_{pred} - N\bar{Y}}{N\bar{Y}}\right] \times 100 \tag{17}$$

ii. **Relative Root Mean Square Error**

$$\%\ RRMSE = \left[\frac{1}{S}\sum_{j=1}^{S}\left(\frac{\hat{T}_{pred} - N\bar{Y}}{N\bar{Y}}\right)^2\right]^{1/2} \times 100 \tag{18}$$

where, S represents number of sampling iterations.

7. Results and Discussion

Table 1 presents the simulation results based on % RB of the proposed estimator of population total and other traditional estimators H-T estimator, Ratio estimator and regression estimator. RRMSE of these estimators are presented in Table 2. Figs. 2 and 3 present pictorial representation of the comparison of the proposed estimator with respect to other traditional estimators of population total on the basis of absolute %RB and RRMSE respectively.

From Table 1 and Fig. 1, it can be seen that the prediction based proposed estimator of the population total under Two-phase sampling gives negligible bias. From Table 2, it can be seen that RRMSE of the proposed prediction based estimator is lowest in comparison to other traditional estimators. So, it can be concluded that the proposed prediction based estimator under Two-phase sampling performs well.

8. Conclusion

In the present study, an estimator based on design and prediction approach has been proposed for the population total in the context of two-phase sampling design. Statistical performance of the proposed estimator has been studied empirically with the help of a simulation study based on various measures like %RB and RRMSE. From simulation results, it can be seen

that the prediction based proposed estimator for the population total is showing better performance than the traditionally existing estimators under two-phase sampling.

Acknowledgement

The authors are very much thankful to the Chief Editor and anonymous referee for their valuable comments which led to improvement of this article.

References

Ahuja, T.K. and P. Misra (2020). A generalized double sampling estimator of population mean using auxiliary information in survey sampling. *International Journal of Agricultural and Statistical Sciences*, **16(2)**, 751-756.

Bandyopadhyay, A. and G.N. Singh (2016). Predictive estimation of population mean in two-phase sampling. *Communications in Statistics-Theory and Methods*, **45(14)**, 4249-4267.

Choudhury, S. and B.K. Singh (2018). A general class of estimators for finite population mean using auxiliary variable in double sampling. *International Journal of Agricultural and Statistical Sciences*, **14(1)**, 351-360.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Edition, New York: John Wiley & Sons.

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology*, **27(2)**, 143-154.

Hidiroglou, M.A. and C.E. Sarndal (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, **24**, 11-20.

Horvitz, D.G. and D.J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

Kadilar, C. and H.Cingi (2004). Ratio estimators in simple random sampling. *Applied Mathematics and Computation*, **151(3)**, 893-902.

Misra, S., S.K. Kumar, D.K. Yadav and A.K. Shukla (2017). An improved ratio type predictive estimator for estimating finite population mean using auxiliary information. *International Journal of Engineering Sciences & Research Technology*, **6(6)**, 524-30.

Misra, P. (2018). Improved double sampling estimator of population mean using auxiliary information. *International Journal of Agricultural and Statistical Sciences*, **14(1)**, 181-186.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, **33**, 101-116.

Panda, K.B. (2015). An efficient predictive approach to estimation in Two-phase sampling, International Organisation of Scientific Research. *Journal of Mathematics*, **11(3)**, 78-81.

Royal, R.M. (1970). On Finite Population Sampling theory under certain linear regression models. *Biometrika*, 57(2), 377-87.

Royal, R.M. (1978). *An Empirical Study of Prediction Theory in Finite Population Sampling: Simple Random Sampling and the Ratio Estimator*. *Survey Sampling and Measurement.* Acedamic Press.

Sarndal, C.E., B. Swensson and J.H. Wretman (1992). *Model Assisted Survey Sampling,* New York: Springer-Verlag.

Subzar, M., T.A. Raja, S.Maqbool, J. Rishu and M. Shabeer (2017). Enhancing the ratio type estimators for estimating population mean using auxiliary information in survey sampling. *International Journal of Agricultural and Statistical Sciences*, **13(1)**, 181-186.

Valliant, R., A.H. Dorfman and R.M. Royall (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley, New York.