Optimal Allocation in Domains Mean Estimation Using Double Sampling with Non-Linear Cost Function in the Presence of Non-Response

Alilah David Anekeya¹, *, Ouma Christopher Onyango², Nyongesa Kennedy¹

¹Departments of Mathematics, Masinde Muliro University of Science and Technology, Kakamega, Kenya
²Departments of Statistics and Actuarial Science, Kenyatta University, Nairobi, Kenya

Email address: aliladavid2010@gmail.com (A. D. Anekeya), chrisouma2004@yahoo.com (O. C. Onyango), knyongesa@mmust.ac.ke (N. Kennedy)

*Corresponding author

To cite this article: Alilah David Anekeya, Ouma Christopher Onyango, Nyongesa Kennedy. Optimal Allocation in Domains Mean Estimation Using Double Sampling with Non-Linear Cost Function in the Presence of Non-Response. American Journal of Theoretical and Applied Statistics. Vol. 7, No. 2, 2018, pp. 45-57. doi: 10.11648/j.ajtas.20180702.11

Received: December 13, 2017; Accepted: January 5, 2018; Published: February 12, 2018

Abstract: Studies have been carried out on domain mean estimation using non-linear cost function. However, little has been done on domain stratum estimation using non-linear cost function using ratio estimation in the presence of non-response. This study develops a method of optimal stratum sample size allocation in domain mean estimation using double sampling with non-linear cost function in the presence of non-response. To obtain an optimum sample size, Lagrangian multiplier technique is employed by minimizing precision at a specified cost. In the estimation of the domain mean, auxiliary variable information in which the study and auxiliary variables both suffer from non-response in the second phase sampling is used. The expressions of the biases and mean square errors of the proposed estimator have also been obtained.

Keywords: Optimal Allocation, Double Sampling, Non-Linear Cost Function, Non-Response

1. Introduction
1.1. Domains

In sampling, estimates are made in each of the class into which the population is subdivided. Such subgroups or classes are known as the domain of study. Units of domains may sometimes be identified prior to sampling. Such domains are called planned domains. For unplanned domain the units cannot be identified prior to sampling and hence the estimates of certain domains is often evident only after the sampling design has been identified or after the sampling and field work have been completed. Hence the size of unplanned domain cannot be controlled. The sample sizes for sub-populations are random variables since formation of these sub-populations is unrelated to sampling.

According to Eurostat [4] the precision threshold and or minimum effective sample sizes are set up for effective planned domains. The minimum sample sizes required to achieve a relative margin error of 100.k% for the total

\[
Y_d = \sum_{i=1}^{n} y_{di}
\]

(Domain total) of a study variable \(y\) over domain \(U\) of size \(N\) given by

\[
n_d = \frac{Z_{1-\alpha/2}^2 N^2 S^2}{K^2 Y^2 + Z_{1-\alpha/2}^2 N^2 S^2}
\]

Where \(S^2\) is the variance of \(Y\) over the domain and \(Z_{1-\alpha/2}\) is the percentile value at \(100\% \alpha\%\) of normal distribution with mean 0 and variance 1, \(K\) is the relative margin of error expressed as a proportion while 100.k% is the relative margin error expressed as a percentage. The population value \(Y_d\) and \(S^2\) is unknowns and have to be estimated using data from auxiliary sources.

1.2. Optimal Allocation with Non-Linear Cost Function

Optimal sample allocation involves determining the
sample size \(n_1, n_2, ..., n_H\) that minimizes the various cost characters under a given sampling budget \(C\) (where \(C\) is the upper limit of the total cost of the survey). Linear cost function is appropriate when the cost involved is associated with non-travel activities of survey e.g drawing sample, preparing survey methods, locating, identifying, interviewing respondents and coding data. Such a linear cost function can be of the form, 

\[ C = c' n' + \sum_{h=1}^{H} c_h n_h \]

where 
- \(c'\) = cost of classification per unit 
- \(n'\) = size of the first sample 
- \(c_h\) = Cost of measuring a unit in stratum \(h\) 
- \(n_h\) = number of units in stratum \(h\).

Generally the above linear cost function is mostly applicable when the major cost item is that of taking the measurements on each unit without considering the cost of the distance between the sample units.

Chernyak [1] proposed optimum allocation in double sampling for stratification with a non-linear cost function. The proposed non-linear cost function is of the form, 

\[ C = c' \left(n'\right)^\alpha + \sum_{k=1}^{K} c_k n_k \]

where, \(c'\) is the cost of classification per unit and \(c_k\) is the cost of measuring a unit in stratum \(k\). Another non-linear cost function proposed was logarithmic in nature of the form, 

\[ C = c' \log n' + n \sum_{k=1}^{K} c_k W_k \]

Okafor and Lee [9] employed the double sampling method to estimate the mean of the auxiliary variable and proceeded to estimate the mean of the study variable in a similar way as Cochran [3]. In this method double sampling ratio and regression estimation was considered. The distribution of the auxiliary information was not known and hence the first phase sample was used to estimate the population distribution of the auxiliary variable while the second phase was used to obtain the required information on the variable of the interest. The optimum sampling fraction was derived for the estimators at a fixed cost. Performance of the proposed estimators was compared with those of Hansen and Hurwitz [5] estimators without considering the cost. It was noted that for the results for which cost component was not considered, regression estimator functions were consistent than the Hansen and Hurwitz [5] estimator. Tschuprow [11] and Neyman [8] proposed the allocation procedure which minimizes variance of sample mean under a linear cost function of sample size \(s > \sum_{h=1}^{H} n_h\), where \(n_h\) is the size of the stratum.

Neyman [8] used Lagrange multiplier optimization technique to get optimum sample sizes for a single variable under study. Holmberg [6] addressed the problem of compromised allocation in multivariate Stratified sampling by taking into consideration minimization of some of the variances or coefficient of variation of the population parameters and of some of the efficiency losses which may be as a result of increase in the variance due to the use of compromise allocation. Saini [10] developed a method of optimum allocation for multivariate stratified two stage sampling design by using double sampling. In this method the problem of determining optimum allocations was formulated as non-linear programming problem (NLPP) in which each NLPP has a convex objective function under a single linear constraints. The Lagrange multiplier technique was used to solve the formulated NLPPs. Khan et al. [7] proposed a quadratic cost function for allocating sample size in multivariate stratified random sampling in the presence of non-response in which a separate linear regression estimator is used. In this multi-objective

Non-linear integer programming problem, an extended lexicographic goal programming was used for solution purpose and comparison made with individual optimum techniques. It is observed that in the allocation techniques, the extended lexicographic goal programming gives minimum values of coefficient of variation than the individual optimum and goal programming technique. Choudhry [2] considered sample allocation issues in the context of estimating Sub-populations (stratum and domain) means as well as the aggregate population means under stratified simple random sampling. In this method non-linear programming was used to obtain the optimal sample allocation to the strata that minimizes the total sample sizes subject to a specified tolerance on the coefficient of variation of the estimators of strata and population means.

From the previous studies, a number of researchers have considered a linear cost function when estimating domains. In dealing with non-response most of them have considered subsampling while holding to the idea that the response mechanism is deterministic. This paper therefore focuses on the estimation of domain mean using double sampling for ratio estimation with non-linear cost function with a random response mechanism. In this study we therefore establish an efficient and cost effective method of estimating domains when the travel component is inclusive and it is not linear.

2. Estimation of Domain Mean and Variance in the Presence of Non-Response

2.1. Introduction

The problem of non-response is inherent in many surveys. It always persists even after call-backs. The estimates obtained from incomplete data will be biased especially when the respondents are different from the non-respondents. The non-response error is not so important if the characteristics of the non-responding units are similar to those of the responding units. However, such similarity of characteristics between two types of units (responding and non-responding) is not always attainable in practice. In double sampling when the problem of
non-response is present, the strata are virtually divided into two disjoint and exhaustive groups of respondents and non-respondents. A sub-sample from non-responding group is then selected and a second more extensive attempt is made to the group so as to obtain the required information. Hansen and Hurwitz [5] proposed a technique of adjusting the non-response to address the problem of bias. The technique consists of selecting a sub-sample of the non-respondents through specialized efforts so as to obtain an estimate of non-responding units in the population. This sub-sampling procedure albeit costly, it’s free from any assumption hence, one does not have to go for a hundred percent response which can be substantially more expensive.

In developing the concept of domain theory with non-response the following assumptions are made;

i. Both the domain study and auxiliary variables suffers from non-response.

ii. The responding and non-responding units are the same for the study and auxiliary characters.

iii. The information on the domain auxiliary variable \( X_d \) is not known and hence \( X_d \) is not available.

iv. The domain auxiliary variables do not suffer from non-response in the first phase sampling but suffers from non-response in the second phase of sampling.

### 2.2. Proposed Domain Estimators

Let \( U \) be a finite population with \( N \) known first stage units. The finite population is divided into \( D \) domains; \( U_1, U_2, \ldots, U_D \) of sizes \( N_1, N_2, \ldots, N_d, \ldots, N_D \) respectively. Further, let \( U_d \) be the domain constituents of any population size \( N_d \) which is assumed to be large and known. Let \( U \) and \( N \) be defined as, \( U = \bigcup_{d=1}^{D} U_d \) and \( N = \sum_{d=1}^{D} N_d \) respectively.

Let \( Y_d \) and \( X_d \) be the domain study and auxiliary variables respectively. Further, let \( \bar{Y}_d \) and \( \bar{X}_d \) be their respective domain population means and auxiliary means with \( y_{d,i} (i = 1,2,3,\ldots,N_d) \) and \( x_{d,i} (i = 1,2,\ldots,N_d) \) observations on the \( i^{th} \) unit. In estimating the domain auxiliary population mean \( \bar{X}_d \) double sampling design is used.

A large first phase sample of size \( n' \) is selected from \( N \) units of the population by simple random sampling without replacement (SRSWOR) design from which \( n_d' \) out \( n' \) first sample units falling in the \( d^{th} \) domain. The assumption here is that all the \( n' \) units supply information of the auxiliary variable \( X_d \) at first phase. A smaller second phase sample of size \( n \) is selected from \( n' \) by SRSWOR from which \( n_d \) out of \( n \) second phase sample units fall in the \( d^{th} \) domain.

For estimating the domain population mean \( \bar{X}_d \) of the auxiliary variables \( X_d \) from a large first phase sample of size \( n_d' \), values of the observations \( x_{d,i} (i = 1,2,3,\ldots,n_d') \) are obtained and a sample auxiliary domain mean \( \bar{x}_d \) is computed. From the second sample of size \( n_d \), let \( y_d \) and \( x_d \) be the domain study and auxiliary observations with \( (i = 1,2,3,\ldots,n_d) \). Let \( n_d \) units supply the information on \( y_d \) and \( x_d \) respondents while \( n_d \) be the non-respondents for both the study and the auxiliary domain variables respectively such that,

\[
    n_d = n_d + n_d
\]

For the \( n_d \) non-respondent group at the second phase sampling, an SRSWOR of \( r_d \) units is selected with an inverse sampling rate of \( v_d \) such that,

\[
    r_d = \frac{n_d}{v_d}, \text{ With } v_d > 1
\]

All the \( r_d \) units respond after making extra efforts of subsampling \( r_d \) non-responding units. In developing the framework of double sampling there are two strata that are non-overlapping and disjoint. Stratum one consists of those units that would respond in the first attempt of phase two with domain population units \( N_d = N_d - N_d \). Both \( N_d \) and \( N_d \) units are not known in advance. The stratum weights of the responding and non-responding groups are defined by \( w_d = \frac{N_d}{N_d} \) and \( w_d = \frac{N_d}{N_d} \) respectively with their estimators defined by \( \hat{w}_d = \frac{n_d}{n_d} \) and \( \hat{w}_d = \frac{n_d}{n_d} \).

Following the Hansen and Hurwitz [5] techniques, the unbiased estimator for estimating the domain population mean using \( \{n_d + r_d\} \) observations on \( y_d \) domain study character is given by;

\[
    \bar{y}_d = \frac{n_d}{n_d} \bar{y}_d + \frac{n_d}{n_d} \bar{y}_d \]

\[
    = w_d \bar{y}_d + w_d \bar{y}_d
\]

(2)

Similarly the estimate for domain auxiliary variable is given by;

\[
    \bar{x}_d = \frac{n_d}{n_d} \bar{x}_d + \frac{n_d}{n_d} \bar{x}_d
\]

\[
    = w_d \bar{x}_d + w_d \bar{x}_d
\]

(3)

Where \( \bar{y}_d \) and \( \bar{x}_d \) are the sample domain means for the
observation $\bar{y}_d$ and $\bar{x}_d$ respectively.

In estimating the overall domain population mean in the presence of non-response, double sampling ratio estimation of the domain mean is used. Define;

$$\tilde{Y}_{d_1} = \frac{\bar{y}_d}{\bar{x}_d} \bar{x}_d = r_d \bar{x}_d \quad \text{and} \quad \tilde{Y}_{d_2} = \frac{\bar{y}_d}{\bar{x}_d} \bar{x}_d = r_d \bar{x}_d$$

With the assumption that,

$$E[\tilde{x}_d] = E[\bar{x}_d] = \bar{x}_d \quad \text{and} \quad E(\bar{y}_d) = \bar{y}_d \quad \text{2.3}$$

3. Mean Square Error of the Ratio Estimator

The expression for the Mean square error (MSE) of $\tilde{Y}_{d_1}$ is derived by the use of the Taylor’s series approximation.

$$\epsilon_{d_1} = \frac{\bar{y}_d - \bar{y}_d}{\bar{y}_d} \Rightarrow \bar{y}_d = \bar{y}_d \left( \epsilon_{d_1} + 1 \right)$$

$$\epsilon_{d_2} = \frac{\bar{y}_d - \bar{x}_d}{\bar{x}_d} \Rightarrow \bar{x}_d = \bar{x}_d \left( \epsilon_{d_2} + 1 \right)$$

With the assumption that $E(\epsilon_{d_1}) = E(\epsilon_{d_2}) = E(\epsilon_{d_3}) = 0$

Further define;

$$\epsilon_{d_3} = \frac{\bar{y}_d - \bar{x}_d}{\bar{x}_d}$$

$$\text{Where,}$$

$$S^2_{\tilde{x}_d} = \text{Variance of the whole domain population mean of the study variable} \ X_d$$

$$S^2_{\tilde{x}_{d_2}} = \text{Variance of the domain population mean for the stratum of non-respondents for Stratum of non-respondents for the study variable} \ Y_d$$

Consider also

$$E[\epsilon^2_{d_3}] = E\left[\frac{\bar{y}_d - \bar{x}_d}{\bar{x}_d} \right] = \frac{1}{\bar{x}_d^2} E\left[\bar{x}_d - \bar{x}_d \right]^2$$

$$= \frac{1}{\bar{x}_d^2} \text{Var}(\bar{x}_d)$$

and $\bar{y}_{d_3}$ are derived by the use of the Taylor’s series approximation.
Next consider

\[ E\left( e_d^2 \right) = E\left[ \frac{\bar{x}_d - \bar{X}_d}{X_d} \right]^2 = \frac{1}{X_d^2} E\left[ \bar{x}_d - \bar{X}_d \right]^2 = \frac{1}{X_d^2} \text{Var}\left( \bar{x}_d \right) \]

Consider,

\[ E\left[ e_d e_d \right] = E\left[ \left( \frac{\bar{y}_d - \bar{Y}_d}{\bar{Y}_d} \right) \left( \frac{\bar{x}_d - \bar{X}_d}{X_d} \right) \right] = \frac{1}{X_d Y_d} E\left[ \left( \bar{y}_d - \bar{Y}_d \right) \left( \bar{x}_d - \bar{X}_d \right) \right] = \frac{1}{X_d Y_d} \text{Cov}\left( \bar{y}_d, \bar{x}_d \right) \]

\[ = \frac{1}{X_d Y_d} \text{Cov}\left[ E\left( \frac{\bar{y}_d}{n_d} \right) E\left( \frac{\bar{x}_d}{n_d} \right) \right] + \frac{1}{X_d Y_d} E\left[ \text{Cov}\left( \frac{\bar{y}_d}{n_d}, \bar{x}_d \right) / n_d \right] + \frac{1}{X_d Y_d} E\left[ \text{Cov}\left( \bar{y}_d, \frac{\bar{x}_d}{n_d} \right) / n_d \right] \]

\[ = \frac{1}{X_d Y_d} E\left[ \text{Cov}\left( \frac{\bar{y}_d}{n_d}, \bar{x}_d \right) / n_d \right] + \frac{1}{X_d Y_d} E\left[ \text{Cov}\left( \bar{y}_d, \frac{\bar{x}_d}{n_d} \right) / n_d \right] \]

\[ = \left( \frac{1}{n_d} - \frac{1}{N_d} \right) \rho_{x_d y_d} \frac{S_{x_d}}{\bar{X}_d} \frac{S_{y_d}}{\bar{Y}_d} + \left( \frac{v_d - 1}{n_d} \right) \rho_{x_d y_d} \frac{S_{x_d}}{\bar{X}_d} \frac{S_{y_d}}{\bar{Y}_d} \]

\[ = \left( \frac{1}{n_d} - \frac{1}{N_d} \right) \rho_{x_d y_d} \frac{S_{x_d}}{\bar{X}_d} \frac{S_{y_d}}{\bar{Y}_d} + \left( \frac{v_d - 1}{n_d} \right) \rho_{x_d y_d} \frac{S_{x_d}}{\bar{X}_d} \frac{S_{y_d}}{\bar{Y}_d} \] (9)

Next,

\[ E\left[ e_d e_d \right] = \left( \frac{\bar{y}_d - \bar{Y}_d}{\bar{Y}_d} \right) \left( \frac{\bar{x}_d - \bar{X}_d}{X_d} \right) = \frac{1}{X_d Y_d} E\left[ \left( \bar{y}_d - \bar{Y}_d \right) \left( \bar{x}_d - \bar{X}_d \right) \right] = \frac{1}{X_d Y_d} \text{Cov}\left( \bar{y}_d, \bar{x}_d \right) \]

\[ = \frac{1}{X_d Y_d} \text{Cov}\left[ E\left( \frac{\bar{y}_d}{n_d} \right) E\left( \frac{\bar{x}_d}{n_d} \right) \right] + \frac{1}{X_d Y_d} E\left[ \text{Cov}\left( \frac{\bar{y}_d}{n_d}, \bar{x}_d \right) / n_d \right] + \frac{1}{X_d Y_d} E\left[ \text{Cov}\left( \bar{y}_d, \frac{\bar{x}_d}{n_d} \right) / n_d \right] \]

\[ = \frac{1}{X_d Y_d} E\left[ \text{Cov}\left( \frac{\bar{y}_d}{n_d}, \bar{x}_d \right) / n_d \right] + \frac{1}{X_d Y_d} E\left[ \text{Cov}\left( \bar{y}_d, \frac{\bar{x}_d}{n_d} \right) / n_d \right] \]

\[ = \left( \frac{1}{n_d} - \frac{1}{N_d} \right) \rho_{x_d y_d} \frac{S_{x_d}}{\bar{X}_d} \frac{S_{y_d}}{\bar{Y}_d} \]

\[ = \left( \frac{1}{n_d} - \frac{1}{N_d} \right) \rho_{x_d y_d} \frac{S_{x_d}}{\bar{X}_d} \frac{S_{y_d}}{\bar{Y}_d} \] (10)

Consider,

\[ E\left[ e_d e_d \right] = \left( \frac{\bar{y}_d - \bar{Y}_d}{\bar{Y}_d} \right) \left( \frac{\bar{x}_d - \bar{X}_d}{X_d} \right) = \frac{1}{X_d Y_d} E\left[ \left( \bar{y}_d - \bar{Y}_d \right) \left( \bar{x}_d - \bar{X}_d \right) \right] = \frac{1}{X_d Y_d} \text{Cov}\left( \bar{y}_d, \bar{x}_d \right) \]

\[ = \frac{1}{X_d Y_d} E\left[ \text{Cov}\left( \bar{y}_d, \bar{x}_d \right) / n_d \right] = \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S_{x_d}^2 / X_d \]

(11)

3.1. Mean Square Error (MSE) of the Ratio Estimator \( \hat{y}_{d_n} \)

and \( \hat{y}_{d_2} \) with the Sample Size Allocation

The ratio estimator of \( \hat{y}_{d_n} \) and \( \hat{y}_{d_2} \) can be defined as;
\hat{\bar{Y}}_{d_{hi}} = \frac{\bar{Y}_d \cdot \bar{x}_d}{\bar{Y}_d} = r_d \bar{x}_d \text{ and } \hat{\bar{Y}}_{d_{i}} = \frac{\bar{Y}_d \cdot \bar{x}_d}{\bar{Y}_d} = r_d \bar{x}_d \text{ respectively}

Proposition 1
The mean square error (MSE) of the estimator defined by
\hat{\bar{Y}}_{d_{hi}} = \frac{\bar{Y}_d \cdot \bar{x}_d}{\bar{Y}_d} \text{ is given by;}

\left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{y_d} + \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{d_{hi}} + \left( \frac{v_{d_{i}} - 1}{n_d} \right) W_{d_i} S^2_{d_{i}}

Where,

\begin{align*}
S^2_{d_{hi}} &= S^2_{y_d} + S^2_{x_d} R^2_d - 2 \rho_{x_d y_d} R_d S_{x_d} S_{y_d} \\
S^2_{d_{i}} &= S^2_{y_d} + S^2_{x_d} R^2_d - 2 \rho_{x_d y_d} R_d S_{x_d} S_{y_d}
\end{align*}

With the notations defined as:

\begin{align*}
S^2_{y_d} &= \text{Variance of the whole domain population mean of the study variable } Y_d \\
S^2_{x_d} &= \text{Variance of the domain population mean for the stratum of non-respondents for the study variable } Y_d \\
S^2_{x_d} &= \text{Variance of the whole domain population mean of the auxiliary variable } X_d \\
v_{d_{i}} &= \text{The inverse sampling rate} \\
R_d &= \text{Population ratio of } \bar{Y}_d \text{ to } \bar{X}_d
\end{align*}

Proof
By definition,

\begin{align*}
\text{MSE} (\hat{\bar{Y}}_{d_{hi}}) &= E \left[ \left( \frac{\bar{Y}_d \cdot \bar{x}_d}{\bar{Y}_d} - \bar{Y}_d \right)^2 \right] \\
&= E \left[ \bar{Y}_d \left( 1 + \varepsilon_{d_{i}} \right) \bar{x}_d \left( 1 + \varepsilon_{d_{i}} \right) \right] \\
&= \bar{Y}_d E \left[ \left( 1 + \varepsilon_{d_{i}} \right) \left( 1 + \varepsilon_{d_{i}} \right) \right] \\
&= \bar{Y}_d E \left[ (\varepsilon_{d_{i}} + \varepsilon_{d_{i}} + \varepsilon_{d_{i}} - \varepsilon_{d_{i}}) \left( 1 - \varepsilon_{d_{i}} + \varepsilon_{d_{i}} + ... \right) \right] \\
&= \bar{Y}_d E \left[ (\varepsilon_{d_{i}} + \varepsilon_{d_{i}} - \varepsilon_{d_{i}})^2 \right] \\
&= \bar{Y}_d E \left[ (\varepsilon_{d_{i}})^2 \right] + E \left[ (\varepsilon_{d_{i}})^2 \right] + 2E \left[ (\varepsilon_{d_{i}}) \varepsilon_{d_{i}} \right] - 2E \left[ (\varepsilon_{d_{i}}) \varepsilon_{d_{i}} \right] - 2E \left[ (\varepsilon_{d_{i}}) \varepsilon_{d_{i}} \right]
\end{align*}

\begin{align*}
\text{MSE} (\hat{\bar{Y}}_{d_{hi}}) &= \bar{Y}_d \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{y_d} + \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{d_{hi}} + \left( \frac{v_{d_{i}} - 1}{n_d} \right) W_{d_i} S^2_{d_{i}} \\
&+ \bar{X}_d \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{x_d} + \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{x_d} + \left( \frac{v_{d_{i}} - 1}{n_d} \right) W_{d_i} S^2_{x_d} \\
&+ 2 \left( \frac{1}{n_d} - \frac{1}{N_d} \right) \rho_{x_d y_d} S_{x_d} S_{y_d} \bar{Y}_d - 2 \left( \frac{1}{n_d} - \frac{1}{N_d} \right) \rho_{x_d y_d} S_{x_d} S_{y_d} \bar{Y}_d \\
&- 2 \left( \frac{v_{d_{i}} - 1}{n_d} \right) \rho_{x_d y_d} W_{d_i} S_{x_d} S_{y_d} \bar{Y}_d - 2 \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{y_d} \\
&= \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{y_d} + \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{y_d} + \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{y_d} R^2_d
\end{align*}
\[ -\left( \frac{1}{n_d} - \frac{1}{N_d} \right) S_{xy}^2 + \left( \frac{v_{d_2} - 1}{n_d} \right) W_{d_2} \left[ S_{xy}^2 + R_d^2 S_{x_j}^2 - 2 \rho_{x_j y_j} R_d S_{x_j} S_{y_j} \right] \]

\[ +2 \left( \frac{1}{n_d} - \frac{1}{N_d} \right) \rho_{x_j y_j} R_d S_{x_j} S_{y_j} - 2 \left( \frac{1}{n_d} - \frac{1}{N_d} \right) \rho_{x_j y_j} R_d S_{x_j} S_{y_j} \]

\[ = \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S_{xy}^2 + \left( \frac{1}{n_d} - \frac{1}{n_d} \right) \left[ S_{xy}^2 + S_{xy}^2 R_d^2 - 2 \rho_{x_j y_j} R_d S_{x_j} S_{y_j} \right] \]

\[ + \left( \frac{v_{d_2} - 1}{n_d} \right) W_{d_2} \left[ S_{xy}^2 + R_d^2 S_{x_j}^2 - 2 \rho_{x_j y_j} R_d S_{x_j} S_{y_j} \right] \]

\[ = \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S_{xy}^2 + \left( \frac{1}{n_d} - \frac{1}{n_d} \right) S_{x_j y_j}^2 + \left( \frac{v_{d_2} - 1}{n_d} \right) W_{d_2} S_{x_j y_j}^2 \]

**Proposition 2**

The mean square error (MSE) of the ratio estimator \( \hat{Y}_{d_2} \) is given by:

\[ \left( \frac{v_{d_2} - 1}{n_d} \right) W_{d_2} S_{x_j y_j}^2 \]

Where,

\[ S_{x_j y_j}^2 = S_{xy}^2 + S_{xy}^2 R_d^2 + 2 \rho_{x_j y_j} R_d S_{x_j} S_{y_j} \]

\[ S_{x_j y_j}^2 = S_{xy}^2 + S_{xy}^2 R_d^2 + 2 \rho_{x_j y_j} R_d S_{x_j} S_{y_j} \]

With the notations as defined in proposition 1 above

**Proof**

MSE of \( \hat{Y}_{d_2} \) = \( \text{MSE} \left( \hat{Y}_{d_2} \right) = E \left[ \hat{Y}_{d_2} - \bar{Y}_d \right]^2 \]

\[ = E \left[ \frac{\bar{Y}_d}{\bar{x}_d} \bar{x}_d - \bar{Y}_d \right]^2 \]

\[ = E \left[ \bar{Y}_d \left( \frac{1 + \epsilon_{d_0}}{1 + \epsilon_{d_1}} \right)^2 \right] \]

\[ = \bar{Y}_d^2 E \left[ \left( \epsilon_{d_1} + \epsilon_{d_0} + \epsilon_{d_1} \epsilon_{d_0} - \epsilon_{d_1} \right) \left( 1 - \epsilon_{d_1} + \epsilon_{d_1}^2 + \ldots \right) \right] \]

\[ = \bar{Y}_d^2 E \left( \epsilon_{d_1} + \epsilon_{d_0} - \epsilon_{d_1} \right)^2 \]

\[ = \bar{Y}_d^2 \left[ E \left( \epsilon_{d_1}^2 \right) + E \left( \epsilon_{d_1} \right)^2 + E \left( \epsilon_{d_1} \right) + 2E \left( \epsilon_{d_1} \epsilon_{d_0} \right) - 2E \left( \epsilon_{d_1} \epsilon_{d_0} \right) - 2 \left( \epsilon_{d_1} \epsilon_{d_2} \right) \right] \]

\[ = \bar{Y}_d^2 \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S_{xy}^2 + \left( \frac{1}{n_d} - \frac{1}{n_d} \right) S_{x_j y_j}^2 + \left( \frac{v_{d_2} - 1}{n_d} \right) W_{d_2} S_{x_j y_j}^2 \]
3.2. Optimal Allocation in Double Sampling for Domain Estimation

An optimum size of a sample is required so as to balance the precision and cost involved in the survey. The optimum allocation of a sample size is attained either by minimizing the precision against a given cost or minimizing cost against a given precision. In this study, a non-linear cost function has been considered.

Denote the cost function for the ratio estimation by

\[ C_d = c_d^\prime \left(n_d^\prime \right)^\theta + c_{d_1}n_d + c_{d_2}n_{d_1} + c_{d_2}n_{d_1} \]  

(12)

Where,

\(c_d^\prime\) = The cost of measuring a unit in the first sample of size \(n_d^\prime\).

\(c_{d_1}\) = The cost of measuring a unit of the first attempt on \(y_d\) with second phase sample size \(n_d\).

\(c_{d_2}\) = The cost associated with the sub-sample of size \(n_{d_1}\) from non-respondents of size \(n_d\).

However the first sample of size \(n_d\) and sub-samples of size \(n_d\) are not known until the first attempt is carried out. The cost will therefore be used in the planning for the survey. Hence the expected cost values of sizes \(n_d\) and \(n_{d_1}\) will be given by: \(n_d = W_d n_d\) and \(n_{d_1} = W_d n_d / \nu_d\). Hence the expected cost function is;

\[ E\left[C_d\right] = c_d^\prime \left(n_d^\prime \right)^\theta + c_{d_1}n_d + c_{d_2}n_{d_1} + c_{d_2}W_d n_{d_1} \]
\[
C_d^* = c_d^*(n_d^*)^\theta + n_d\left(c_{d_1} + c_{d_2}W_{d_1} + c_{d_2}\cdot \frac{W_{d_2}}{v_{d_2}}\right) \tag{13}
\]

3.3. Results of Double Sampling for Domain Estimation in the Presence of Non-Response

Proposition 3

The variance for the estimated domain mean for the estimated domain mean \(\hat{Y}_{d_1} = \frac{\bar{Y}_d}{\bar{x}_d}\) is minimum for a specified cost \(C_d^*\) when, \(n_d^* = \left(\frac{S_{d_1}^2 \cdot \theta c_d^*}{\lambda}\right)^{\frac{1}{\theta + 1}}\)

Next the partial derivative with respect to \(v_{d_2}\) obtained as;

\[
\frac{\partial G(W_d)}{\partial v_{d_2}} = \frac{W_{d_2}S_{d_{b_2}}^2}{n_d} - \frac{\lambda c_{d_1}W_d n_d}{v_{d_2}^2} = 0
\]

\[
W_{d_2}S_{d_{b_2}}^2 = \lambda c_{d_1}W_d n_d
\]

\[
v_{d_2} = \frac{n_d\sqrt{\lambda c_{d_1}}}{S_{d_{b_2}}}
\]

Consider the equation

\[
G(W_d) = \left(\frac{1}{n_d^*} - \frac{1}{N_d}\right)S_{y_d}^2 + \left(\frac{1}{n_d^*} - \frac{1}{n_d}\right)S_{b_1}^2 + \left(\frac{v_{d_2} - 1}{n_d}\right)W_{d_2}S_{d_{b_2}}^2 + \lambda\left[c_d^*(n_d^*)^\theta + n_d\left(c_{d_1} + c_{d_2}W_{d_1} + c_{d_2}\cdot \frac{W_{d_2}}{v_{d_2}}\right)\right]
\]

\[
\frac{\partial G(W_d)}{\partial n_d^*} = -\frac{S_{y_d}^2 + S_{b_1}^2}{n_d^*} + \frac{S_{d_{b_2}}^2}{n_d^*} + \lambda c_d^*(n_d^*)^{-1} = 0
\]

\[
\frac{\partial G(W_d)}{\partial v_{d_2}} = 0
\]

\[
\frac{n_d}{v_{d_2}} = \frac{S_{d_{b_2}}}{\lambda c_{d_1}}
\]

Substituting this in Equations (17) we obtain into (16) we obtain
The partial derivative of the equation (18) with respect to \( n_d \) is obtained as

\[
\frac{\partial G(W_d)}{\partial n_d} = -\frac{S_{d_h}^2}{n_d^2} + \frac{W_d S_{d_h}^2}{n_d^2} + \lambda \left( c_{d_0} + c_{d_1} W_{d_1} \right) = 0
\]

\[
= -S_{d_h}^2 + W_d S_{d_h}^2 + \lambda n_d^2 \left( c_{d_0} + c_{d_1} W_{d_1} \right) = 0
\]

\[
\lambda n_d^2 \left( c_{d_0} + c_{d_1} W_{d_1} \right) = S_{d_h}^2 - W_d S_{d_h}^2
\]

\[
n_d^2 = \frac{S_{d_h}^2 - W_d S_{d_h}^2}{\lambda \left( c_{d_0} + c_{d_1} W_{d_1} \right)}
\]

\[
n_d = \sqrt{\frac{S_{d_h}^2 - W_d S_{d_h}^2}{c_{d_0} + c_{d_1} W_{d_1}}}
\]

Where \( \varnothing = \frac{1}{\lambda} \). But \( v_d = \frac{S_{d_h}}{S_{d_h}} n_d \) from equation (17)

Then,

\[
C_d^* = c_d' \left( \frac{S_{d_h}^2 \theta}{\theta + 1} \right) + 1 \sqrt{\frac{S_{d_h}^2 - W_d S_{d_h}^2}{c_{d_0} + c_{d_1} W_{d_1}}} \left( c_{d_0} + c_{d_1} W_{d_1} + c_{d_2} W_{d_2} \sqrt{c_{d_2}} \right)
\]

\[
C_d^* = c_d' \left( \frac{S_{d_h}^2 \theta}{\theta + 1} \right) + 1 \sqrt{\frac{S_{d_h}^2 - W_d S_{d_h}^2}{c_{d_0} + c_{d_1} W_{d_1}}} \left( c_{d_0} + c_{d_1} W_{d_1} + c_{d_2} W_{d_2} \sqrt{c_{d_2}} \right)
\]

Let,

\[
A = \left( c_d' \frac{S_{d_h}^2 \theta}{\theta + 1} \right)^{\frac{\theta}{\theta + 1}}
\]

\[
B = \left( c_{d_0} + c_{d_1} W_{d_1} \right) \left( \frac{S_{d_h}^2 - W_d S_{d_h}^2}{c_{d_0} + c_{d_1} W_{d_1}} \right)^{\theta} + W_d S_{d_h} \sqrt{c_{d_2}}
\]

\[
C = C_d^*
\]

The equation (19) becomes;

\[
A \frac{\theta}{\theta + 1} + B \lambda^{\frac{1}{2}} - C = 0
\]

If \( \theta = 1 \) and substituting this value in the equation (20) we obtain a linear equation of the form
\[ A 2^{1/2} + B \lambda^{1/2} - C = 0 \]

With the values of \( A \) and \( B \) defined as:

\[ A = \left( c_d' \right)^{1/2} \left( \frac{S_{ey}^2}{\theta} \right)^{1/2}, \quad B = \left( c_d + c_d W_{d_d} \right) \sqrt{\frac{S_{dy}^2 - W_{d_d} S_{dy}^2}{c_d + c_d W_{d_d}}} + W_{d_d} S_{dy} \sqrt{c_d} \] and \( C = C_d^* \)

Solving the linear equation solution obtained is, \[ \lambda = \sqrt{\frac{A + B}{C}} \]

When \( \theta = \frac{1}{3} \) and substituting this value in the equation (20) we obtain a linear equation of the form,

\[ A \lambda^{1/2} + B \lambda^{1/2} - C = 0 \]

With the values of \( A \) defined as;

\[ A = \left( c_d' \right)^{1/2} \left( \frac{S_{ey}^2}{\theta} \right)^{1/2} \]

While \( B \) and \( C \) remains as earlier defined

Solving the equation (21) solution obtained is,

\[ \lambda = \left[ \frac{2A}{\sqrt{B^2 + 4AC}} - B \right]^{1/2} \]

Proposition 4

If the expected cost function is of the form \( c_d' \)

\[ c_d' \log n_d' + n_d \left( c_d + c_d W_{d_d} + c_d \frac{W_{d_d}}{v_{d_d}} \right) \] then the variance of the

\[ G(W_d) = \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S_{zy}^2 + \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S_{dy}^2 + \left( \frac{v_{d_d} - 1}{n_d} \right) W_{d_d} S_{dy}^2 \]

\[ + \lambda \left[ c_d' \log n_d' + n_d \left( c_d + c_d W_{d_d} + c_d \frac{W_{d_d}}{v_{d_d}} \right) - C_d^* \right] \] (22)

To obtain the normal equations for the expression (22) the equation is differentiated partially with respect to \( n_d' \) and the partial derivatives are equated to zero

\[ \frac{\partial G(W_d)}{\partial n_d'} = -\frac{S_{zy}^2}{n_d' n_d} - \frac{S_{dy}^2}{n_d' n_d} + \lambda c_d' n_d' = 0 \]

\[ = -S_{zy}^2 + S_{dy}^2 + \lambda c_d' n_d' = 0 \]

\[ \lambda c_d' n_d' = S_{zy}^2 - S_{dy}^2 \]
But, \( S^2_{\alpha_n} = S^2_{y_d} - S^2_{d_n} > 0 \), thus,

\[
n'_d = \frac{S^2_{\alpha_n}}{\lambda c'_d}
\]

To solve for \( \lambda \), let the variance be given as \( V_0 \) then substitute the values of \( v_{d_5}, n_d \) and \( n'_d \) into the equation,

\[
G(W_d) = \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{y_d} + \left( \frac{1}{n_d} - \frac{1}{n'_d} \right) S^2_{d_n} + \left( \frac{v_{d_5} - 1}{n_d} \right) W_d S^2_{d_n} = V_0
\]

\[
= \left( \frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{y_d} + \left( \frac{1}{n_d} - \frac{1}{n'_d} \right) S^2_{d_n} + \left( \frac{v_{d_5} - 1}{n_d} \right) W_d S^2_{d_n} - V_0 = 0
\]

\[
= \frac{1}{n_d} \left( S^2_{y_d} - S^2_{d_n} \right) + \frac{1}{n'_d} \left( S^2_{d_n} - W_d S^2_{d_n} \right) + v_{d_5} W_d S^2_{d_n} = \frac{S^2_{y_d}}{N_d} + V_0 = 0
\] (23)

Substitute the values of \( v_{d_5}, n_d \) and \( n'_d \) into the equation (23) and simplify to obtain,

\[
\lambda c'_d + \sqrt{\lambda} \left[ \sqrt{\left( S^2_{d_n} - W_d S^2_{d_n} \right) \left( c_{d_6} + c_{d_5} W_d \right) } \right] + W_d S^2_{d_n} \sqrt{c_{d_5}} = V_0
\]

(24)

Let,

\[
V^*_d = \left( \frac{S^2_{d_n}}{N_d} + V_0 \right), \quad A = c'_d \text{ and } B = \sqrt{\left( S^2_{d_n} - W_d S^2_{d_n} \right) \left( c_{d_6} + c_{d_5} W_d \right) + W_d S^2_{d_n} \sqrt{c_{d_5}}}
\]

Thus equation (24) becomes,

\[
A\lambda + B \lambda^{3/2} - V^*_d = 0
\] (25)

Solving for \( \lambda \) in equation (21) the solution becomes,

\[
\lambda = \left[ \frac{\sqrt{B^2 + 4AC} - B}{2A} \right]^{3/2}
\]

4. Conclusion

It is noted that value of inverse sampling rate \( v_{d_5} \) do not depend on the value of the Lagrangian multiplier \( \lambda \).

Further the value of sampling rate, \( v_{d_5} < 1 \), if \( c_{d_5} \) (the unit cost associated with the sub-sample of size \( r_{d_5} \) from non-respondents of size \( n_{d_5} \) ) is less than both \( c_{d_6} \) (the unit cost for processing the responded data of \( y_d \) at the first attempt of size \( n_d \) ) and \( c_{d_5} \) (the cost of measuring a unit of the first attempt on \( y_d \) with second phase sample size \( n_d \) ) and also when the value of \( S^2_{d_n} \) is not too large relative to \( S^2_{d_n} \). The second phase sample size \( n_d \) will be minimum if the value of \( \left( S^2_{d_n} - W_d S^2_{d_n} \right) > 0 \) but less than 1. If the value of \( S^2_{d_n} = S^2_{\alpha_n} - S^2_{d_n} < 1 \) and \( S^2_{d_n} > S^2_{d_n} \) with the values of \( c_{d_5} \) (the cost of measuring a unit in the first sample of size \( n_d \) ) not being too large to the relative \( S^2_{d_n} \) then the value of \( n'_d \) (size of the first sample) will be minimum. These minimum values therefore make the theoretic cost survey of the proposed estimator as minimal as possible.
References

[1] Cherniyak O. I., (2001). Optimal allocation in stratified sampling and double sampling with non-linear cost function, *Journal of Mathematical Sciences* 103, 4 pp. 525-528.

[2] Choudhry H. G., Rao, J. N. K, and Michael A., Hidiroglou, (2012). On sample allocation for efficient domain estimation, *Survey methodology*, 38 (1) pp. 23-29.

[3] Cochran W. G., (1977) *Sampling techniques*. New York: John Wiley and Sons, (1977).

[4] Eurostat., (2008). Introduction to Sample Design and Estimation Techniques, *Survey Sampling Reference Guidelines*. Luxembourg; Office for Publication of the European Communities pp. 36.

[5] Hansen M. H. and Hurwitz W. W, (1946). The problem of non-response in sample surveys. *The Journal of the American Statistical Association*, 41 517-529.

[6] Holmberg A., (2002). A multi-parameter perspective on the choice of sampling designs in surveys. *Journal of statistics in transition*. 5 (6) pp. 969-994.

[7] Khan S. U., Muhammad Y. S., and Afgan N., (2009). Multi-objective compromise allocation stratified sampling in the presence of non-response using quadratic cost function. *International Journal of Business and social science*. 5 (13) pp. 162-169.

[8] Neyman, C. and Jerzy D. (1934); On the Two Different Aspects of the Representative methods of stratified sampling and the method of purposive selection, *Journal of royal statistical society*. 97 (4) pp. 558-625.

[9] Okafor F. C, (2001). Treatment of non-response in successive sampling, *Statistica*, 61 (2) 195-204.

[10] Saini M., and Kumar A, (2015). Method of Optimum allocation for Multivariate Stratified two stage Sampling design Using double Sampling. *Journal of Probability and Statistics forum*, 8, pp. 19-23.

[11] Tschuprow and Al A., (1923). On mathematical expectation of the moments of frequency distribution in the case of correlated observation (chapters 4-6) *Metron* 2 (1) pp. 646-683, (1923).