STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 369–388

ESTIMATING POPULATION MEAN WITH MISSING DATA IN UNEQUAL PROBABILITY SAMPLING

Kajal Dihidar [1]

ABSTRACT

Nonresponse problem is a serious obstacle to the validity of estimates in a survey. The estimates become biased due to the missing values in data. The problem is how to deal with missing values, once they have been deemed impossible to recover. One way of exploring a possible lack of representativity in missing data is to estimate the response probabilities which are usually done by logistic regression model. However, the drawback of the logit model is that this requires values of the explanatory variables of the model to be known for all nonrespondents. Bethlehem (2012) showed that the response probabilities can be estimated by some weighting adjustment technique without having the individual data of the nonrespondents. Here we consider the doubtful nature of nonresponse regarding possible existence of relationship with any of the covariates. Moreover, instead of simple random sampling, we consider general unequal probability sampling scheme for selecting respondents. This paper presents the modification of Bethlehem (2012) proposal for unequal probability sampling to obtain the unbiased estimators for population total/average of a variable of interest and variance estimator and compares them with the usual estimators through numerical simulations.

Key words: non-response, missing at random, missing completely at random, unequal probability sampling.

1. Introduction

Almost all large scale sample surveys suffer the problem with missing data. It may occur even if an investigator tries to have all questions fully responded to in a survey, or if the respondent is not available at home to answer the questionnaire. One of the effects of nonresponse is that the sample size is smaller than expected. This would lead to less accurate, but still valid estimates of population characteristics, which can be taken care of by taking the initial sample size larger. A far more serious effect of nonresponse is that estimates of population characteristics may be biased. This situation occurs if, due to nonresponse, some groups in the population are over- or under-represented, and these groups behave differently with respect

[1] Sampling and Official Statistics Unit, Indian Statistical Institute, Kolkata, West Bengal, India. E-mail: kajaldihidar@gmail.com, dkajal@isical.ac.in.

to the characteristics to be investigated. Consequently, wrong conclusions will be drawn from the survey data. The amount of bias created due to missing values often increases with the rate of occurence of nonresponse. Above all, the large number of missing values in the data set can also lead to computational difficulties.

Towards this problem, starting from the pioneered work by Hansen and Hurwitz (1946), many methods of attempts to re-collect the missing values in sample surveys are available in the literature. However, in most of the practical sample survey work, it is not possible to recover the actual missing values. In such situations, the problem is how to estimate the population parameters dealing with the missing values. The method of response modeling and imputation are popular to survey statisticians in this direction. Good details regarding this are given in Rubin (1987) and S$\breve{a}$rndal, Swenson and Wretman (1992).

In general, obtaining the responses from the selected units is totally unknown in advance. For this reason, the probabilistic models are assumed to describe the unknown response distributions. Politz and Simmons (1949, 1950) obtained the response probability of a respondent as the proportion of time staying at home. The response probability may be directly related to the study variable and hence to the auxiliary variable, which is highly related to the study variable. For example, in the study of household income, the people with high income may respond with low probability and may be under represented in the sample. Similarly, if tax return is considered as an auxiliary variable, then the response probability of an individual may be inversely proportional to the amount of tax return.

Regarding the possible relatioship of missingness with any of the covariates, Rubin (1976) defined the concepts of missing at random (MAR) and missing completely at random (MCAR). Missing completely at random (MCAR) means that the missing data is not related to the values of any variable, neither to the response variable itself nor to other covariates, whether missing or observed; whereas missing at random (MAR) means that the missing data is unrelated to the actual missing values but is related either to observed covariates or to observed response variable itself or to both. Among many contributors in this area, Folsom (1991), Fuller et al. (1994), Kott (2006), Chang and Kott (2008) and Kott and Chang (2010) advocated the use of calibration weighting to adjust for unit nonresponse. In this regard, for more detailed clarification, interested researchers may see Heitzan and Basu (1996), Singh (2010).

In case the covariate relation is considered, the concept of the response propensity is introduced in Little (1986). The response propensity is the probability of response given the values of some auxiliary variables. The response propensities are also unknown, so they need to be estimated. For this purpose, the logistic model is used in practice. Of course, another model sometimes used is the probit model. Estimates of the coefficients in both the logit and probit models are obtained by maximum likelihood estimation. And the estimated response propensities in these two models are always in the interval [0, 1]. However, the drawback of the logit and probit models is that these require the values of the explanatory variables of the model to be known for all nonrespondents. Bethlehem (2012) showed that this condition can be relaxed by computing response probabilities from weights that have been obtained from some weighting adjustment technique. This technique produces weights that correct for the lack of representativity of the survey response. Since the weights can be seen as a kind of inverted response probabilities, they can be used to estimate response probabilities. Weights are computed following those techniques without having the individual data of the nonrespondents. We use this approach to estimate the response propensities from correlated auxiliary variables.

In this paper, we consider the situation where some of the respondents selected using an unequal probability sampling scheme fail to respond and the nature of nonresponse is uncertain as to whether it is MAR or MCAR. Moreover, instead of considering the simple random sampling, we consider any general unequal probability sampling scheme even without replacement for selecting the respondents because we believe that many of the practical cases of large-scale sample surveys require the selection of respondents with probability proportional to size measures of some auxiliary variable related to study variable. Under the consideration of doubtful nature of random nonresponse, we shall derive here unbiased estimators for population total/average of a variable of interest and variance estimators in unequal probability sampling scheme. The derived estimators will be compared with usual estimators in presence of random nonresponse through numerical simulations.

We organize our findings in the following sections.

2. Unbiased estimator of population mean and variance with missing data

Suppose in a finite survey population $U = (1, \ldots, i, \ldots, N)$ a person labelled i has the value y_i defined on a variable y of interest and has value $x_i > 0$ defined on an auxiliary variable x closely related to the study variable y. The values of x are all positive and known for all the population units in U. Our problem is to estimate $\bar{Y} = \dfrac{1}{N}\sum_{i=1}^{N} y_i$ on the basis of a sample s of size n, selected with probability $p(s)$ according to a sampling design p.

Let π_i and π_{ij} be the first and second order inclusion probabilities of the units in U. Let us define a random variable δ_i as

$$\delta_i = \begin{cases} 1 & \text{if } i^{th} \text{ unit responds,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Let E_P, V_P denote the expectation and variance operators with respect to the sampling design for selecting the respondents. Let E_R, V_R denote the expectation and variance operators with respect to obtaining a response from the selected respondent, and E, V denote the overall expectation and variance operators. In this setup, δ_i is a Bernoulli random variable with probability of success as δ_i^*, say, and it is known. So, $E_R(\delta_i)$ = Prob$(\delta_i = 1)$ = δ_i^*, and $V_R(\delta_i)$ = $\delta_i^*(1 - \delta_i^*)$. We first of all assume that the value of response probability depends on some auxiliary variables which are well correlated with the study variable, but the exact relationship of the response probability with the auxiliary variables is unknown to us. We can get possible relationships based on some statistical testing whether the available data is MCAR or MAR. For example, we can apply some variable selection method to see whether or not the response propensity depends (or not) on a set of auxiliary variables. However, a weighted sum of these two estimators (MCAR and MAR) would be an alternative to choosing one over the other to balance their degree of bias. This type of estimator, namely the 'composite estimator' is formed by compromising in between the MAR estimator and MCAR estimator, with a compromising factor $\lambda (0 < \lambda < 1)$.

The composite estimator of population total Y will be obtained as

$$\hat{Y}_{comp} = \lambda \hat{Y}_{MCAR} + (1 - \lambda)\hat{Y}_{MAR},$$

where $\hat{Y}_{MCAR}$ and $\hat{Y}_{MAR}$ respectively denote the MCAR and MAR estimators for Y. We may get the optimal compromising factor by minimmizing the MSE of the composite estimator with respect to λ under the assumption that the covariance factor of $\hat{Y}_{MCAR}$ and $\hat{Y}_{MAR}$ is too small relative to the MSE of $\hat{Y}_{MAR}$ and then it can be negligible. In this situation, the optimal compromising factor λ_{opt} may be obtained as

$$\lambda_{opt} = \frac{MSE(\hat{Y}_{MAR})}{MSE(\hat{Y}_{MCAR}) + MSE(\hat{Y}_{MAR})}.$$

In practical situation, λ_{opt} can be estimated by substituting the estimates of $MSE(\hat{Y}_{MCAR})$ and of $MSE(\hat{Y}_{MAR})$ based on the sample survey data in above expression of λ_{opt}.

2.1. Unbiased estimator of population mean

Under the non-response setup, a homogeneous linear unbiased estimator for population mean is

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{i \in s} y_i b_{si}\left(\frac{\delta_i}{\delta_i^*}\right) = \frac{1}{N}\sum_{i \in s} u_i b_{si}, \text{ where } u_i = y_i \frac{\delta_i}{\delta_i^*} \tag{2}$$

and b_{si}'s are free of y_i's and satisfy $\sum_{s \ni i} p(s) b_{si} = 1, \forall\, i \in U$.

This happens because

$$E_R(u_i) = \frac{y_i}{\delta_i^*} E_R(\delta_i) = \frac{y_i}{\delta_i^*}\delta_i^* = y_i, \tag{3}$$

and

$$E\left(\hat{\bar{Y}}\right) = E_P E_R\left[\frac{1}{N}\sum_{i \in s} u_i b_{si}\right] = E_P\left[\frac{1}{N}\sum_{i \in s} b_{si} E_R(u_i)\right]$$
$$= E_P\left[\frac{1}{N}\sum_{i \in s} y_i b_{si}\right] = \bar{Y}. \tag{4}$$

2.2. Variance of the unbiased estimator of population mean

From the definition of u_i, we have

$$V_R(u_i) = \frac{y_i^2}{\delta_i^{*2}} V_R(\delta_i) = \frac{y_i^2(1-\delta_i^*)}{\delta_i^*}. \tag{5}$$

So, the variance of the estimator given in Eqn. (5) is

$$\begin{aligned} V\left[\hat{\bar{Y}}\right] &= V_P E_R\left[\frac{1}{N}\sum_{i\in s} u_i b_{si}\right] + E_P V_R\left[\frac{1}{N}\sum_{i\in s} u_i b_{si}\right] \\ &= V_P\left[\frac{1}{N}\sum_{i\in s} y_i b_{si}\right] + E_P\left[\frac{1}{N^2}\sum_{i\in s} b_{si}^2 V_R(u_i)\right] \\ &= \frac{1}{N^2}\left[\sum_{i=1}^N y_i^2 c_i + \sum_{i=1}^N \sum_{j=1, j\neq i}^N y_i y_j c_{ij} + E_P\left(\sum_{i\in s} b_{si}^2 \frac{y_i^2}{\delta_i^*}(1-\delta_i^*)\right)\right] \\ &= \frac{1}{N^2}\left[\sum_{i=1}^N y_i^2 c_i + \sum_{i=1}^N \sum_{j=1, j\neq i}^N y_i y_j c_{ij} + \left(\sum_{i=1}^N \frac{y_i^2 b_{si}}{\delta_i^*}(1-\delta_i^*)\right)\right], \end{aligned} \tag{6}$$

where $c_i = E_P(b_{si}^2 I_{si}) - 1$ and $c_{ij} = E_P(b_{si} b_{sj} I_{sij}) - 1$ where I_{si} and I_{sij} are defined as

$$I_{si} = \begin{cases} 1 & \text{if } i \in s, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

and $I_{sij} = I_{si} I_{sj}$.

2.3. Unbiased variance estimator for population mean

First of all, we find an unbiased estimator for $V_R(u_i)$. We note that $\delta_i^2 = \delta_i$ and so,

$$E_R(u_i^2) = E_R\left[\frac{y_i^2 \delta_i^2}{\delta_i^{*2}}\right] = E_R\left[\frac{y_i^2 \delta_i}{\delta_i^{*2}}\right] = \frac{y_i^2}{\delta_i^{*2}} E_R[\delta_i] = \frac{y_i^2}{\delta_i^*},$$

and so

$$E_R[u_i^2 \delta_i^*] = y_i^2. \tag{8}$$

Now,

$$\begin{aligned} V_R(u_i) &= E_R(u_i^2) - (E_R(u_i))^2 = E_R(u_i^2) - y_i^2 \\ &= E_R(u_i^2) - E_R[u_i^2 \delta_i^*] = E_R[u_i^2(1-\delta_i^*)] \end{aligned} \tag{9}$$

implies that

$$v_R(u_i) = \hat{V}_R(u_i) = u_i^2(1-\delta_i^*). \tag{10}$$

Let c_{si} and c_{sij} be such that $E_P(c_{si}I_{si}) = c_i$ and $E_P(c_{sij}I_{sij}) = c_{ij}$.

We define

$$v_1 = \frac{1}{N^2}\left[\sum_{i\in s} u_i^2 c_{si} + \sum_{i\in s}\sum_{j\in s, j\neq i} u_i u_j c_{sij} + \sum_{i\in s} v_R(u_i)(b_{si}^2 - c_{si})\right], \tag{11}$$

and

$$v_2 = \frac{1}{N^2}\left[\sum_{i\in s} u_i^2 c_{si} + \sum_{i\in s}\sum_{j\in s, j\neq i} u_i u_j c_{sij} + \sum_{i\in s} v_R(u_i) b_{si}\right]. \tag{12}$$

Following Raj (1966), we have $E_P E_R(v_1) = V(\hat{\hat{Y}}) = E_P E_R(v_2)$, and so v_1 and v_2 are two unbiased estimators for $V(\hat{\hat{Y}})$.

3. Estimation of response probability

The true response probability δ_i^* as discussed in Section 2 is practically unknown in advance. So, we need to use an estimator for this.

If no covariate relation is considered, the missing data is considered as missing completely at random (MCAR), then the probability of response (assuming same for all units) is estimated by $\frac{r}{n}$, where n is the sample size and r is the number of responses obtained out of n persons sampled.

If the covariate relation is considered, the concept of the response propensity is introduced in Little (1986). He has defined the response propensity of element i as

$$\delta_i^*(\mathbf{X}) = P(\delta_i = 1|\mathbf{X}_i), \tag{13}$$

where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ is a vector of values of, say, p auxiliary variables. So, the response propensity is the probability of response given the values of some auxiliary variables. The response propensities are also unknown, so they need to be estimated.

3.1. Traditional models

The most frequently used model to estimate the response propensities is the logistic regression model. It assumes the relationship between response propensity and auxiliary variables as

$$\text{logit}(\delta_i^*(\mathbf{X})) = \log\left(\frac{\delta_i^*(\mathbf{X})}{1-\delta_i^*(\mathbf{X})}\right) = \sum_{j=1}^{p} X_{ij}\beta_j, \tag{14}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a vector of p regression coefficients.

Of course, another model, the probit model can also be used. It assumes the relationship between response propensity and auxiliary variables as

$$\text{probit}(\delta_i^*(\mathbf{X})) = \Phi^{-1}\left(\delta_i^*(\mathbf{X})\right) = \sum_{j=1}^{p} X_{ij}\beta_j, \tag{15}$$

where Φ^{-1} is the inverse of the $N(0,1)$ distribution function.

Estimates of the coefficients in both the logit and probit models can be obtained by maximum likelihood estimation. And the estimated response propensities in these two models are always in the interval [0, 1].

However, the drawback of the logit and probit models is that these require the values of the explanatory variables of the model to be known for all nonrespondents. But this is not the situation in many cases. To overcome this drawback, we follow Bethlehem (2012) model to estimate the response propensities and this is described below.

3.2. Bethlehem Model

Bethlehem (2012) showed how to estimate the response probabilities from weights that have been obtained from some weighting adjustment technique without having the individual data of the nonrespondents. The basic idea is to assign weights to responding elements in such a way that over-represented groups get a weight smaller than 1 and under-represented groups get a weight larger than 1. There is a relationship between response probabilities and weights: large weights correspond to small response probabilities, and vice versa. Therefore, it should be possible to transform weights into estimates for response probabilities.

There are several types of weighting techniques. The most frequently used ones are post-stratification, generalized regression estimation and raking ratio estimation. Weighting is based on the use of auxiliary information. Auxiliary information is defined here as a set of variables that have been measured in the survey, and for which the distribution in the population, or in the complete sample, is available. The individual values of the auxiliary variables are not required for the nonresponding elements. Among several weighting techniques, we adopt here the generalized regression estimation technique for simplicity. The generalized regression estimator is based on a linear model that attempts to explain a target variable of the survey from one or more auxiliary variables. The weights resulting from generalized regression estimation make the response representative with respect to the auxiliary variables in the model (S$\ddot{a}$rndal, 2011).

In principle, the auxiliary variables in the linear model have to be continuous variables, i.e. they measure a size, value or duration. However, it is also possible to use categorical variables. The trick is to replace a categorical variable by a set of dummy variables, where each dummy variable represents a category, i.e. it indicates whether or not a person belongs to a specific category. Suppose there are p

(continuous) auxiliary variables available. The p-vector of values of these variables for element i is $\mathbf{X}_i$.

Let Y be the N-vector of all values in the population of the target variable, and let $\mathbf{X}$ be the $N \times p$-matrix of all values of the auxiliary variables. The vector of population means of the p auxiliary variables is defined by

$$\bar{\mathbf{X}} = (\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2, \ldots, \bar{\mathbf{X}}_p)'.$$

We assume that this vector representing the population information is available, based on some expert guess or on the result of some prior survey. If the auxiliary variables are correlated with the target variable, then for a suitably chosen vector $\mathbf{B} = (B_1, B_2, \ldots, B_p)'$ of regression coefficients for a best fit of Y on $\mathbf{X}$, the residuals $E = (E_1, E_2, \ldots, E_N)'$, defined by $E = Y - \mathbf{XB}$ will vary less than the values of the target variable itself. The population regression coefficient B obtained by applying ordinary least squares technique is

$$B = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \left(\sum_{i=1}^{N} \mathbf{X}_i\mathbf{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i\mathbf{Y}_i\right). \tag{16}$$

The vector $\mathbf{B}$ can be estimated by

$$\mathbf{b} = \left(\sum_{i\in s} \pi_i^{-1}\mathbf{x}_i\mathbf{x}_i'\delta_i\right)^{-1} \left(\sum_{i\in s} \pi_i^{-1}\mathbf{x}_i\mathbf{y}_i\delta_i\right), \tag{17}$$

where π_i is the first order inclusion probability of unit i in sample s.

Let $\mathbf{X}_s$, $\mathbf{Y}_s$ be $n \times p$ and $n \times 1$ versions of $\mathbf{X}$ and $\mathbf{Y}$ for the units $i \in s$ where n is the sample size. Let $\mathbf{W}_s$ be the $n \times n$ diagonal matrix with the weights w_i for the units $i \in s$ on the diagonal. The Horvitz-Thompson (1952) weights are $w_i = 1/\pi_i$. Also let δ_s be the $n \times n$ diagonal matrix with values δ_i for the units $i \in s$ on the diagonal. The vector $\mathbf{b}$ can then be written in matrix form as

$$\mathbf{b} = (\mathbf{X_s'W_s}\delta_\mathbf{s}\mathbf{X_s})^{\mathbf{-1}}(\mathbf{X_s'W_s}\delta_\mathbf{s}\mathbf{Y_s}). \tag{18}$$

The generalized regression estimator is now defined by

$$\bar{y}_{GR} = \frac{1}{N}\left[\sum_{i\in s}\frac{y_i\delta_i}{\pi_i} + (\mathbf{X} - \sum_{i\in s}\pi_i^{-1}\mathbf{x_i}\delta_\mathbf{i})'\mathbf{b}\right]. \tag{19}$$

Following Bethlehem and Keller (1987), the generalized regression estimator can be rewritten in the form of the weighted estimator as

$$\bar{y}_{GR} = \sum_{i\in s} w_i y_i \delta_i, \tag{20}$$

where the weights are

$$w_i = \bar{\mathbf{X}}\left[\sum_{j\in s}\pi_j^{-1}\mathbf{x}_j'\mathbf{x}_j\delta_j\right]^{-1}\pi_i^{-1}\mathbf{x}_i', \tag{21}$$

where $\mathbf{x}_j$ is the k-dimensional vector of control variables, $\bar{\mathbf{X}}$ is the row vector of population totals of the control variables, the first element of $\mathbf{x}_j$ is always one, and the first element of $\bar{\mathbf{X}}$ is one.

Following Bethlehem (2012), we get the adjusted weight w_i for observed element i for unequal probability sampling, as equal to $w_i = \nu' \mathbf{X}_i$, where ν is a vector of weight coefficients defined by

$$\nu = (\sum_{i \in s} \pi_i^{-1} \delta_i) \left(\sum_{j \in s} \pi_j^{-1} \mathbf{x}_j \mathbf{x}_j' \delta_j \right)^{-1} \bar{X}. \tag{22}$$

So, it is clear that computation of the weight does not require the individual values of the nonresponding elements. It is sufficient to have the population means of the auxiliary variables.

As an illustration, the case of one auxiliary variable with C categories is considered. Then C dummy variables $X^{(1)}$,$X^{(2)}$, ..., $X^{(C)}$ are defined. For an observation in a category H, the corresponding dummy variable is assigned the value 1, and all other dummy variables are set to 0. Consequently, the vector of population means of these dummy variables is equal to

$$\bar{X} = \left(\frac{N_1}{N}, \frac{N_2}{N}, ..., \frac{N_C}{N} \right), \tag{23}$$

where N_j is the number of elements in category j (in the population), for $j = 1, 2, ..., C$. The vector ν of weight coefficients is equal to

$$\nu = \frac{\sum_{i \in s} \pi_i^{-1} \delta_i}{N} \left(\frac{N_1}{\sum_{i \in s} \pi_i^{-1} \delta_i X^{(1)}}, \frac{N_2}{\sum_{i \in s} \pi_i^{-1} \delta_i X^{(2)}}, ..., \frac{N_C}{\sum_{i \in s} \pi_i^{-1} \delta_i X^{(C)}}, \right)'. \tag{24}$$

Now we see how the weights computed by means of generalized regression estimation can be transformed into response propensities. Let there be p categorical auxiliary variables. The continuous variables can also be transformed into categorical variables by forming several meaningful groups. The values of these variables for unit i are denoted by the vector

$$X_i = (X_i^{(1)}, X_i^{(2)}, ..., X_i^{(p)})'.$$

The number of categories of variable $X^{(j)}$ is denoted by C_j, say, for $j = 1, 2, ..., p$. So, for variable $X^{(j)}$, the categories are numbered as $1, 2, ..., C_j$.

We note from the above adjusted regression weight formula that all responding units with the same set of values for the auxiliary variables will be assigned the same weight. Suppose a unit is in category number k_1 of the first variable, category k_2 of the second variable,..., and category k_p of the p^{th} variable. Let $w(k_1, k_2, ..., k_p)$ denote the corresponding weight. Furthermore, we assume that there are $r(k_1, k_2, ..., k_p)$ respondents in this group. The number of sample units

$n(k_1, k_2, ..., k_p)$ in the group can now be estimated by

$$\hat{n}(k_1, k_2, ..., k_p) = \frac{\sum_{i \in s} \pi_i^{-1}}{\sum_{i \in s} \pi_i^{-1} \delta_i} \times w(k_1, k_2, ..., k_p) \times r(k_1, k_2, ..., k_p). \quad (25)$$

The response propensity for all elements in the group can be estimated by

$$\hat{\rho}(k_1, k_2, ..., k_p) = \frac{r(k_1, k_2, ..., k_p)}{\hat{n}(k_1, k_2, ..., k_p)} = \frac{\sum_{i \in s} \pi_i^{-1} \delta_i}{\sum_{i \in s} \pi_i^{-1}} \times \frac{1}{w(k_1, k_2, ..., k_p)}. \quad (26)$$

So, it is clear that the response propensities are inversely proportional to the weights. We note that the response propensities can only be estimated for respondents and not for nonrespondents.

Following Chaudhuri (2010), we can now obtain several competitive estimators and variance estimators for population mean of a variable of interest by replacing the response probabilities δ_i^* with their estimates $\hat{\delta}_i^*$ obtained by whatever means using MCAR or the logit/probit models or the $\hat{\rho}(k_1, k_2, ..., k_p)$s of Bethlehem model in the respective equations shown in Section 2.

4. Illustrative simulation based findings

In this section, we present the results of numerical comparison of our different estimators based on sample drawn using unequal probability sampling scheme. To perform the comparison simulation, we use the data of a real population. The population considered is the Labor Force Population obtained from the September 1976 Current Population Survey (CPS) conducted in the United States and this data set was studied by Valliant et al. (2000). This population data contains information on demographic and economic variables from the persons chosen in that labor force survey. This is basically a clustered population of individuals, where the clusters are compact geographic areas used as one of the stages of sampling in the CPS and are typically composed of about four nearby households. The units within clusters for this illustrative population are individual persons. For our numerical illustration, we use all of the observations of one stratum containing information of $N = 210$ persons. This data set contains information of persons about their usual number of hours of working per week, usual amount of their weekly wages along with their demographic and social charateristics like their age, sex, race (non-black, black). We consider the usual amount of their weekly wages as the main variable y of interest and the usual number of hours of working per week as the size measure variable x for drawing sample of persons. Our objective is to estimate the average weekly wage taking into account the doubtful missing information obtained from the selected respondents chosen by varying probability sampling scheme and to study the performance behaviour of alternative estimators. We use the logistic model as $\phi(x_i) = \frac{1}{1+e^{(-1.65+.5\times race_i+.08\times sex_i+0.05\times age_i)}}$ to generate the true probabilities δ_i^* s.

4.1. Application in two specific unequal probability sampling schemes

For illustration in practical sample survey situation, we consider two different unequal probability sampling schemes. The first one is Midzuno's (1952) scheme and the second one is a modification of Brewer's (1963) scheme. The choice of these two different types of sampling schemes is based on the knowledge of having a constant effective sample size and uniformly non-negative variance estimator for Midzuno's scheme and the knowledge of having varying effective sample size and uniformly non-negative variance estimator for the modified Brewer's scheme. We now describe briefly these two sampling schemes.

4.1.1. Midzuno's scheme

Midzuno (1952) suggested this scheme first by drawing one unit by probability proportional to the size measure of an auxiliary variable with known $x_i > 0$, for $i = 1, 2, \ldots, N$. Then, keeping the selected unit aside, the remaining $(n-1)$ units should be chosen by simple random sampling without replacement (SRSWOR) out of $(N-1)$ units. Let $X = \sum_{i=1}^{N} x_i$. Then, under this scheme,

$$\pi_i = \frac{x_i}{X} + \frac{X - x_i}{X} \frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} = \frac{x_i}{X} \frac{N-n}{N-1} + \frac{n-1}{N-1} \;\; \forall i = 1, 2, \ldots, N, \tag{27}$$

and

$$\begin{aligned} \pi_{ij} &= \frac{x_i}{X} \frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} + \frac{x_j}{X} \frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} + \frac{X - x_i - x_j}{X} \frac{\binom{N-3}{n-3}}{\binom{N-1}{n-1}} \\ &= \frac{x_i + x_j}{X} \frac{(N-n)(n-1)}{(N-1)(N-2)} + \frac{(n-1)(n-2)}{(N-1)(N-2)}, \end{aligned} \tag{28}$$

$\forall i \neq j \in U$. For this scheme, $\pi_i \pi_j > \pi_{ij} \forall i \neq j \in U$, and so for the Horvitz and Thompson (1952)'s estimator $\sum_{i \in s} \frac{y_i}{\pi_i}$ for population total Y of y variable, the Yates and Grundy (1953) form of variance estimator $V_{YG} = \sum_{i \in s} \sum_{j \in s, j > i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$ is always non-negative.

Now, keeping in mind that all the y_i's may not be available for all $i \in s$, so with respect to the response probabilities δ_i^*, an unbiased estimator for population mean is

$$e_M = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i} \left(\frac{\delta_i}{\delta_i^*} \right) = \frac{1}{N} \sum_{i \in s} \frac{u_i}{\pi_i}, \text{ where } u_i = y_i \frac{\delta_i}{\delta_i^*}, \tag{29}$$

since $E_R(u_i) = y_i$ and $E_P E_R(e_M) = E_P(\frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}) = \bar{Y}$.

The variance of e_M is obtained as

$$V(e_M) = V_P E_R(e_M) + E_P V_R(e_M) = \frac{1}{N^2}\left[V_P\left(\sum_{i\in s}\frac{y_i}{\pi_i}\right) + E_P\left(\sum_{i\in s}\frac{1}{\pi_i^2}V_R(u_i)\right)\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + E_P\left(\sum_{i\in s}\frac{1}{\pi_i^2}\frac{y_i^2(1-\delta_{ik}^*)}{\delta_i^*}\right)\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{1}{\pi_i}\frac{y_i^2(1-\delta_i^*)}{\delta_i^*}\right]. \quad (30)$$

Following Chaudhuri, Adhikary and Dihidar (2000), an unbiased estimator of the variance of e_M is :

$$v(e_M) = \frac{1}{N^2}\left[\sum_{i\in s}\sum_{j\in s,j>i}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{u_i}{\pi_i} - \frac{u_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{v_R(u_i)}{\pi_i}\right]. \quad (31)$$

Next, as δ_i^* is unknown to us, so following Chaudhuri (2010), we can have the required estimators and variance estimators of the population mean as $\hat{e}_M = e_M|\{\delta_i^* = \hat{\delta}_i^*\}$ and $v(e_M)|\{\delta_i^* = \hat{\delta}_i^*\}$, i.e. replacing δ_i^* in e_M and $v(e_M)$ throughout by its estimate $\hat{\delta}_i^*$ obtained by any means as discussed earlier.

4.1.2. Modified Brewer's scheme

We consider the following scheme of Brewer (1963), modified by Seth (1966) and further modified by Chaudhuri and Pal (2002). Let us call the normed size measues of auxiliary variable as $p_i = \frac{x_i}{X}$'s for $i = 1, 2, \ldots, N$. In this scheme, on the first draw, the unit i is chosen with a probability proportional to $q_i = \dfrac{p_i(1-p_i)}{1-2p_i}$ and leaving aside the unit i so chosen, a second unit $j(\neq i)$ is chosen in the second draw from the remaining units with the probability $\dfrac{p_j}{1-p_i}$. Writing $D = \displaystyle\sum_{i=1}^{N}\frac{p_i}{1-2p_i}$, from Brewer (1963) it is known that the inclusion probability of i and that of the pair $(i,j), i \neq j$ in the sample of 2 draws are respectively

$$\pi_i(2) = 2p_i, \quad \text{and} \quad \pi_{ij}(2) = \left[\frac{2p_ip_j}{1+D}\right]\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right). \quad (32)$$

It is further known that

$$\Delta_{ij}(2) = \pi_i(2)\pi_j(2) - \pi_{ij}(2) \geq 0 \;\; \forall i, j(i \neq j) \in U. \quad (33)$$

We use '2' within parenthesis to emphasize that this scheme uses 2 draws. Let the sample chosen as above be augmented by adding to the 2 distinct units so drawn as above, $(r-2)$ further distinct units from the remaining $(N-2)$ units of U by simple random sampling without replacement (SRSWOR). For such a scheme introduced by Seth (1966) admitting r distinct units in each sample, the inclusion probabilities $\pi_i(r)$ for i and $\pi_{ij}(r)$ for $(i,j)(i \neq j)$, involving r draws, are respectively

$$\pi_i(r) = \frac{1}{N-2}\left[(r-2) + (N-r)\pi_i(2)\right], \tag{34}$$

$$\begin{aligned}\pi_{ij}(r) &= \pi_{ij}(2) + \left(\frac{r-2}{N-2}\right)(\pi_i(2) + \pi_j(2) - 2\pi_{ij}(2)) \\ &+ \left(\frac{r-2}{N-2}\right)\left(\frac{r-3}{N-3}\right)(1 - \pi_i(2) - \pi_j(2) + \pi_{ij}(2)).\end{aligned} \tag{35}$$

Chaudhuri and Pal (2002) modified this sampling scheme of Seth (1966) by allowing $(r-2)$ to be (1) a number $(n-2)$ to be chosen with a pre-assigned probability $w(0 < w < 1)$ and (2) a number $(n-1)$ to be chosen with the complementary probability $(1-w)$. Then, a sample s so drawn will have a size n with probability w and $(n+1)$ with probability $(1-w)$. So, the effective sample size is either n or $(n+1)$. So for this modified sampling scheme if π_i^* and π_{ij}^* denote the first and second order inclusion probabilities, then

$$\pi_i^* = w\pi_i(n) + (1-w)\pi_i(n+1), \tag{36}$$

and

$$\pi_{ij}^* = w\pi_{ij}(n) + (1-w)\pi_{ij}(n+1). \tag{37}$$

Chaudhuri and Pal (2002) also showed that $\pi_i^*\pi_j^* \geq \pi_{ij}^*, \forall i, j \in U(i \neq j)$.

Under this scheme, for the Horvitz and Thompson (1952) estimator $\displaystyle\sum_{i\in s}\frac{y_i}{\pi_i^*}$ for population total Y of y variable, the variance estimator is given by Chaudhuri and Pal (2002)

$$v_{CP} = \sum_{i\in s}\sum_{j\in s, j>i}\frac{\pi_i^*\pi_j^* - \pi_{ij}^*}{\pi_{ij}^*}\left(\frac{y_i}{\pi_i^*} - \frac{y_j}{\pi_j^*}\right)^2 + \sum_{i\in s}\frac{\alpha_i y_i^2}{\pi_i^{*2}}, \tag{38}$$

where $\alpha_i = 1 + \dfrac{1}{\pi_i^*}\displaystyle\sum_{j=1,j\neq i}^{N}\pi_{ij}^* - \sum_{i=1}^{N}\pi_i^*$. They also showed that $\alpha_i > 0$ for all $i \in U$ and so v_{CP} is always non-negative.

Now, keeping in mind that all the y_i's may not be available for all $i \in s$, so with respect to the response probability δ_i^*, an unbiased estimator for population mean is

$$e_B = \frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i^*}\left(\frac{\delta_i}{\delta_i^*}\right) = \frac{1}{N}\sum_{i\in s}\frac{u_i}{\pi_i^*}, \text{ where } u_i = y_i\frac{\delta_i}{\delta_i^*}, \tag{39}$$

since $E_R(u_i) = y_i$ and $E_P E_R(e_B) = E_P(\frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i^*}) = \bar{Y}$.

The variance of e_B is obtained as

$$V(e_B) = V_P E_R(e_B) + E_P V_R(e_B)$$

$$= \frac{1}{N^2}\left[V_P\left(\sum_{i\in s}\frac{y_i}{\pi_i^*}\right)+E_P\left(\sum_{i\in s}\frac{1}{\pi_i^{*2}}V_R(u_i)\right)\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i^*\pi_j^*-\pi_{ij}^*)\left(\frac{y_i}{\pi_i^*}-\frac{y_j}{\pi_j^*}\right)^2+\sum_{i=1}^{N}\frac{\alpha_i y_i^2}{\pi_i^*}+E_P\left(\sum_{i\in s}\frac{1}{\pi_i^{*2}}\frac{y_i^2(1-\delta_i^*)}{\delta_i^*}\right)\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i^*\pi_j^*-\pi_{ij}^*)\left(\frac{y_i}{\pi_i^*}-\frac{y_j}{\pi_j^*}\right)^2+\sum_{i=1}^{N}\frac{\alpha_i y_i^2}{\pi_i^*}+\sum_{i=1}^{N}\frac{1}{\pi_i^*}\frac{y_i^2(1-\delta_i^*)}{\delta_i^*}\right]. \tag{40}$$

Following Chaudhuri, Adhikary and Dihidar (2000), an unbiased estimator of the variance of e_B is:

$$v(e_B) = \frac{1}{N^2}\left[\sum_{i\in s}\sum_{j\in s,j>i}\frac{\pi_i^*\pi_j^*-\pi_{ij}^*}{\pi_{ij}^*}\left(\frac{u_i}{\pi_i^*}-\frac{u_j}{\pi_j^*}\right)^2+\sum_{i\in s}\frac{\alpha_i u_i^2}{\pi_i^{*2}}+\sum_{i\in s}\frac{v_R(u_i)}{\pi_i^*}\right]. \tag{41}$$

Next, as δ_i^* is unknown to us, so following Chaudhuri (2010), we can have the required estimators and variance estimators of the population mean as $\hat{e}_B = e_B|\{\delta_i^* = \hat{\delta}_i^*\}$ and $v(e_B)|\{\delta_i^* = \hat{\delta}_i^*\}$ i.e. replacing δ_i^* in e_B and $v(e_B)$ throughout by its estimate $\hat{\delta}_i^*$ obtained by any means as discussed earlier.

4.2. Efficiency comparison

To get some ideas about the estimates and the measure of errors obtained in a practical sample survey situation, we perform the simulation by drawing samples of size equal to 15% of the population size by each of above mentioned sampling schemes taking the usual number of hours of working per week (x) as the size measure. Let us denote the estimators based on the two sampling designs by $\hat{e}_M$ and $\hat{e}_B$, the subscripts M and B being for Midzuno's and Modified Brewer's sampling schemes respectively. The notations used for different competitive estimators for population mean concerned are described as below.

(1) $\hat{e}_M(1)$ and $\hat{e}_B(1)$: Based on the estimate of δ_i^* as $\hat{\delta}_{i1}^* = \frac{r}{n}$, the traditional MCAR estimator.

(2) $\hat{e}_M(2)$ and $\hat{e}_B(2)$: Based on the estimate of δ_i^* as $\hat{\delta}_{i2}^*$ obtained from usual logit model.

(3) $\hat{e}_M(3)$ and $\hat{e}_B(3)$: Based on the estimate of δ_i^* as $\hat{\delta}_{i3}^*$ obtained from Bethlehem (2012) model.

(4) $\hat{e}_M(4, \lambda = ...)$ and $\hat{e}_B(4, \lambda = ...)$: Based on compromising in between traditional MCAR and Bethlehem (2012) model.

We compare the estimators using measures based on confidence intervals for the parameters they are meant to estimate. For each sampling scheme, the sampling is replicated a large number of times, say, 10000 times and the corresponding estimator is computed for each such sample. The standardized pivotal, namely, $\tau = \frac{\hat{\theta}-\theta}{\sqrt{v(\hat{\theta})}}$ is assumed to be a standard normal deviate. Then,

$$\left(\hat{\theta} - 1.96\sqrt{v(\hat{\theta})},\ \ \hat{\theta} + 1.96\sqrt{v(\hat{\theta})}\right)$$

is used as a 95% confidence interval for θ based on the estimator $\hat{\theta}$.

Two measures based on this confidence interval are often used to compare the performance of the alternative estimators. One is the ACP, i.e., the Average Coverage Percentage, which is the percent of the replicated samples for which θ is covered by the above confidence interval. The second measure is the AL, i.e., the average length, which is the length of the confidence interval ($=2 \times 1.96\sqrt{v(\hat{\theta})}$) averaged over all the replicates.

We consider another measure, namely the simulation coefficient of variation (or in short SimCV, say) defined by

$$SimCV(\hat{\theta}) = 100 \times \frac{\sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\hat{\theta}_l - \left(\frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_l\right)\right)^2}}{|\theta|},$$

where

L = the number of replications in the simulation study,

$\hat{\theta}_l$ = the value of the estimator in the l^{th} iteration $(l = 1, 2, \ldots, L)$,

θ = the value of the population parameter computed based on the whole population dataset.

As the simulation CV is not sufficient to compare the accuracies of the estimators, additionally some more values are computed. These are defined below.

(i) Simulation relative biases of the estimators given by:

$$rB(\hat{\theta}) = 100 \times \frac{\frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_l - \theta}{|\theta|},$$

(ii) Simulation relative Root Mean Squared Errors given by:

$$rRMSE(\hat{\theta}) = 100 \times \frac{\frac{1}{L}\sum_{l=1}^{L}\left(\hat{\theta}_l - \theta\right)^2}{|\theta|},$$

(iii) Simulation relative biases of variance estimators given by:

$$rB(\hat{D}^2(\hat{\theta})) = 100 \times \frac{\frac{1}{L}\sum_{l=1}^{L}\hat{D}^2(\hat{\theta}_l) - \hat{D}^2(\hat{\theta})}{\hat{D}^2(\hat{\theta})},$$

where

$\hat{D}^2(\hat{\theta}_l)$ = the value of the variance estimator in the l^{th} iteration $(l = 1, 2, \ldots, L)$, and

$\hat{D}^2(\hat{\theta}) = \frac{1}{L}\sum_{l=1}^{L}\left(\hat{\theta}_l - \left(\frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_l\right)\right)^2.$

A good estimator is the one with a high value of ACP; the closer this value is to 95%, the better the estimator. Again, with respect to AL, a good estimator should have a small value of AL. Similarly, the small values for the criteria SimCV, Simulation relative biases of the estimators, Simulation relative Root Mean Squared Errors, Simulation relative biases of variance estimators are also desirable for a good estimator. We present the results of these comparison criteria in Tables 1 to 4. Almost all the above stated criteria show good performances. More specifically, it is interesting to note that all values of the biases of the estimators are negative, and they are quite small, and the only exception is for the biases of the variance estimators. It is important to note that they are not so quite small, and this inspires us to investigate for other variance estimators in the future research. However, the overall results show that the estimator based on Bethlehem (2012) model used in unequal probability sampling scheme is a good competitor of the traditional estimators. Moreover, the compromised estimators based on the MCAR and Bethlehem (2012) model may also be tried with several compromising factors in order to achieve further improvement.

Table 1. Simulation results for alternative estimators (Midzuno's scheme)

Estimator	$ACP(\hat{\theta})$ (%)	$AL(\hat{\theta})$	$SimCV(\hat{\theta})$ (%)	$rB(\hat{\theta})$	$rRMSE(\hat{\theta})$	$rB(\hat{D^2}(\hat{\theta}))$
MCAR($\hat{e}_M(1)$)	95.8	195.7	14.38	-2.37	14.55	131.99
Logistic($\hat{e}_M(2)$)	97.0	240.1	14.83	-1.82	14.92	136.56
Bethlehem($\hat{e}_M(3)$)	96.3	220.8	13.13	-6.53	14.65	195.54

Table 2. Simulation results for compromised estimators (Midzuno's scheme)

Estimator	$ACP(\hat{\theta})$ (%)	$AL(\hat{\theta})$	$SimCV(\hat{\theta})$ (%)	$rB(\hat{\theta})$	$rRMSE(\hat{\theta})$	$rB(\hat{D^2}(\hat{\theta}))$
$\hat{e}_M(4, \lambda = 0.1)$	95.7	196.5	13.87	-4.31	14.50	134.78
$\hat{e}_M(4, \lambda = 0.2)$	95.2	190.7	13.45	-5.83	14.64	142.56
$\hat{e}_M(4, \lambda = 0.3)$	94.9	190.8	13.11	-7.00	14.84	150.96
$\hat{e}_M(4, \lambda = 0.4)$	95.7	189.2	12.82	-7.86	15.02	159.28
$\hat{e}_M(4, \lambda = 0.5)$	94.8	196.4	12.58	-8.44	15.13	168.10
$\hat{e}_M(4, \lambda = 0.6)$	94.7	199.7	12.40	-8.74	15.16	177.06
$\hat{e}_M(4, \lambda = 0.7)$	95.1	203.2	12.29	-8.76	15.08	185.81
$\hat{e}_M(4, \lambda = 0.8)$	95.4	206.7	12.28	-8.46	14.90	193.31
$\hat{e}_M(4, \lambda = 0.9)$	96.0	219.5	12.46	-7.77	14.67	197.93

Table 3. Simulation results for alternative estimators (Modified Brewer's scheme)

Estimator	$ACP(\hat{\theta})$	$AL(\hat{\theta})$	$SimCV(\hat{\theta})$	$rB(\hat{\theta})$	$rRMSE(\hat{\theta})$	$rB(\hat{D}^2(\hat{\theta}))$
	(%)		(%)			
MCAR($\hat{e}_B(1)$)	96.3	245.23	14.49	-3.14	14.81	137.28
Logistic($\hat{e}_B(2)$)	96.8	263.70	15.43	-1.60	15.50	182.50
Bethlehem($\hat{e}_B(3)$)	98.3	262.08	13.72	-6.46	15.15	219.12

Table 4. Simulation results for compromised estimators (Modified Brewer's scheme)

Estimator	$ACP(\hat{\theta})$	$AL(\hat{\theta})$	(%)	$rB(\hat{\theta})$	$rRMSE(\hat{\theta})$	$rB(\hat{D}^2(\hat{\theta}))$
	(%)		(%)			
$\hat{e}_B(4, \lambda = 0.1)$	97.60	239.21	14.05	-4.82	14.84	145.24
$\hat{e}_B(4, \lambda = 0.2)$	97.40	237.26	13.71	-6.12	15.00	153.24
$\hat{e}_B(4, \lambda = 0.3)$	97.00	236.40	13.43	-7.11	15.19	161.64
$\hat{e}_B(4, \lambda = 0.4)$	98.40	236.53	13.22	-7.82	15.35	170.53
$\hat{e}_B(4, \lambda = 0.5)$	98.00	237.61	13.06	-8.27	15.45	179.91
$\hat{e}_B(4, \lambda = 0.6)$	98.20	239.70	12.96	-8.48	15.48	189.63
$\hat{e}_B(4, \lambda = 0.7)$	98.20	242.89	12.93	-8.43	15.43	199.41
$\hat{e}_B(4, \lambda = 0.8)$	98.20	247.39	13.00	-8.11	15.32	208.66
$\hat{e}_B(4, \lambda = 0.9)$	97.80	253.55	13.22	-7.48	15.18	216.20

5. Concluding remarks

This paper presents a general framework to estimate the population mean in the presence of auxiliary variables and non-response under the unequal probability sampling scheme. It is shown that the good competitive estimators can be obtained by estimating the response probabilities postulating good models keeping in mind that the values of the possible correlated variables may also not be available for the non-respondents. Finally, the doubtful missing data can also be profitably handled with the use of compromised estimator. Moreover, we need to examine the performance of the suggested estimators with some other estimators like Kott and Chang (2010), Chang and Kott (2008). Our research is in progress to see if the results of the proposed estimators in this paper show better performance in comparison with Kott and Chang (2010), Chang and Kott (2008) estimators.

Acknowledgement

The author is indebted to the anonymous referees for many constructive and insightful comments and suggestions which led to an improved version of the manuscript. The author gratefully acknowledges their kind attempt to draw her attention to valuable pioneering research papers in this direction and to inspire her for future research.

REFERENCES

BETHLEHEM, J. G., (2012). Using response probabilities for assessing representativity. *Statistics Netherlands, Discussion Paper.*

BETHLEHEM, J. G., KELLER, W. A., (1987). Linear weighting of sample survey data. *Journal of Official Statistics.* 3, 141-153.

BREWER, K. R. W., (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics.* 5, 5-13.

CHANG, T., KOTT, P. S., (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika.* 95, 557–571.

CHAUDHURI, A., (2010). Essentials of Survey Sampling. PHI Learning Private Limited. New Delhi.

CHAUDHURI, A., ADHIKARY, A., DIHIDAR, S., (2000). Mean square error estimation in multi-stage sampling. *Metrika.* 52, 115-131.

CHAUDHURI, A., PAL, S., (2002). Estimating proportions from unequal probability samples using randomized responses by Warner's and other devices. *Journal of the Indian Society of Agricultural Statistics.* 55(2), 174-183.

FOLSOM, R. E., (1991). Exponential and logistic weight adjustment for sampling and nonresponse error reduction. *Proceedings of Social Statistics Section, Washington, DC: American Statistical Association.* 197-202.

FULLER, W. A., LOUGHIN, M. M., BAKER, H. D., (1994). Regression weighting for the 1987-88 National Food Consumption Survey. *Survey Methodology.* 20, 75-85.

HANSEN, M. H., HURWITZ, W. N., (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association.* 41, 517-529.

HEITJAN, D. F., BASU, S., (1996). Distinguishing 'Missing at Random' and 'Missing Completely at Random'. *The American Statistician.* 50, 207-213.

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

KOTT, P. S., (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology.* 32, 133-142.

KOTT, P. S. CHANG, T., (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association.* 105(491), 1265-1275.

LITTLE, R. J. A., (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review.* 54. 139-157.

MIDZUNO, H., (1952). On the sampling system with probabilities proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics.* 3, 99-107.

POLITZ, A. N., SIMMONS, W. R., (1949). An Attempt to Get 'Not-at-Homes' into the Sample Without Call-Backs. *Journal of the American Statistical Association.* 44, 9-31.

POLITZ, A., SIMMONS, W., (1950). Note on an Attempt to Get 'Not-at-Homes' into the Sample Without Call-Backs. *Journal of the American Statistical Association.* 45, 136-137.

RAJ, D., (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association.* 61, 391-396.

RUBIN, D. B., (1987). Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

RUBIN, D. B., (1976). Inference and missing data. *Biometrika.* 63, 581-592.

SÄRNDAL, C. E., (2011). Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics.* 27, 1-21.

SÄRNDAL, C. E., SWENSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling. Springer-Verlag. New York.

SETH, G. R., (1966). On estimators of variance of estimate of population total in varying probabilities. *Journal of the Indian Society of Agricultural Statistics.* 18, 52-56.

SINGH, S., (2010). Layman's understanding of non-response: How Michael and Amy adjust a missing phone call. *LIAISON, Statistical Society of Canada.* 24(3), p. 67.

VALLIANT, R., DORFMAN, A. H., ROYALL, R. M., (2000). Finite Population Sampling and Inference: A Prediction Approach. Wiley Series in Survey Methodology. New York.

YATES, F., GRUNDY, P. M., (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the American Statistical Association.* 75, 206-211.