International Journal of Probability and Statistics 2014, 3(1): 8-14
DOI: 10.5923/j.ijps.20140301.02

Sample Design for Domain Calibration Estimators

E. P. Clement[1,*], G. A. Udofia[2], E. I. Enang[2]

[1]Department of Mathematics and Statistics, University of Uyo, Uyo, Nigeria
[2]Department of Mathematics, Statistics and Computer Science, University of Calabar, Calabar, Nigeria

Abstract The phenomenon of nonresponse in a sample survey reduces the precision of parameters estimates and increases bias in estimates resulting in larger mean square error, thus ultimately reducing their efficiency. An important technique to address these problems is by calibration. We proposed calibration estimators for totals of domain of study. Sample designs and in particular sample sizes are chosen so as to provide reliable estimates for domains of study. But budget and other constraints usually prevent the allocation of sufficiently large samples to domains to provide reliable estimates using traditional statistical techniques. We have developed an approach for finding the best sample design for the domain calibration estimators subject to a cost constraint and derived optimum stratum sample sizes that minimized the variances of the proposed domain calibration estimators and reduced the objective function. The efficacy of the proposed domain calibration estimators was tested through a real data analysis. Results of the analytical study using real data showed that our proposed domain calibration estimator is substantially superior to the traditional GREG-estimator with relatively small bias, mean square error and average length of confidence interval.

Keywords Design weights, Domain of study, GREG-estimator, Nonresponse, Optimum stratum sample sizes

1. Introduction

Nonresponse always exists when surveying human populations as people hesitate to respond in surveys. Nonresponse as an aspect in almost every type of sample survey creates problems for estimation which cannot simply be eliminated by increasing sample size.

This phenomenon of nonresponse in a sample survey reduces the precision of parameters estimates and increases bias in estimates resulting in larger mean square error, thus ultimately reducing their efficiency.

An important technique to address these problems is by calibration. Calibration as a tool for reweighting for nonresponse was first introduced by [4] for the estimation of finite population characteristics like means ratios and totals.

Deville and Sarndal calibration estimation procedure provides a valuable class of techniques for combining data sources. The basic idea is to use estimates from one set of sources, which may be treated as sufficient accurate to act as benchmark. Estimates based on data from further sample source are then adjusted so as to agree with these benchmarks. The process of adjustment is called Calibration. The constraints that the estimates of the benchmarks based on this source should agree with the benchmarks are called Calibration Constraints.

The problem of calibration of design weights is well known in the literature of survey sampling. [4] used the method of calibration of estimators using auxiliary information. Their calibration method provides a class of estimates. Some of the well known estimators such as classical ratio estimator belong to this class. Several authors including [5-6, 9-13] among others considered the [4] method and derived important calibrated estimators. But so far derivation of calibrated estimators from the class of calibrated estimators derived by [4] method in the context of domain estimation is not well known in the literature. Our objective therefore is to extend calibration estimation to domain estimation.

Consider the finite population under study U of size N divided into D domains; $U_1, U_2, \dots, U_D$ of sizes $N_1, N_2, \dots, N_D$ respectively. Domain membership of any population unit is unknown before sampling. It is assumed that domains are quite large [7].

The technique of estimation by calibration is based on the idea to use auxiliary information to obtain a better estimate of a population statistic. Consider a finite population U of size N with units labels $1,2,\dots,N$. Let $y_k, k = 1,2,\dots,N$ be the study variable and $\boldsymbol{x}_k, k = 1,2,\dots,N$ be the k-dimensional vector of auxiliary variables associated with unit k.

Suppose we are interested in estimating the domain total $Y_d = \sum_{U_d} y_{dk}$. We draw a sample $s = \{1,2,\dots,n\} \in U_d$ using a probability sampling design P, with probability $P(s)$, where the first and second order inclusion probabilities are $\pi_k = P(k \in s)$ and $\pi_{kl} = P(k,l \in s)$ respectively.

An estimate of Y_d is the Horvitz-Thompson (HT)

* Corresponding author:
epclement@yahoo.com (E. P. Clement)
Published online at http://journal.sapub.org/ijps

estimator

$$\hat{Y}_{dHT} = \sum_s d_k y_{dk} \quad (1)$$

where $d_k = 1/\pi_k$ is the sampling weight defined as the inverse of the inclusion probability π_k for unit k.

An attractive property of the HT-estimator is that it is guaranteed to be unbiased regardless of the sampling design P[8]. It variance under P is given as:

$$V_P(\hat{Y}_{HT}) = \sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \quad (2)$$

Suppose there are x_k $\{k = 1,2,\dots,N\}$ auxiliary variables at unit k and $\boldsymbol{x}_k = (x_1,\dots,x_n,\dots,x_N)$ may or may not be known a priori. $\boldsymbol{X}_d = \sum_s x_{dk}$ is the domain total for $\boldsymbol{X}$, and is known a priori. Ideally, we would like

$$\hat{\boldsymbol{X}}_d = \sum_s d_k \boldsymbol{x}_{dk} \quad (3)$$

but often times this is not true.

The idea behind calibration estimation is to find weights $w_k, k = 1,2,\dots,n$ close to d_k based on a distance function such that

$$\hat{\boldsymbol{X}}_{d,w} = \sum_s w_k \boldsymbol{x}_{dk} \quad (4)$$

Equation (4) is the calibration constraint. We wish to find weights w_k similar to d_k so as to preserve the unbiased property of the HT-estimator. Once w_k is found, then our propose calibration estimator for $Y_{d,w}$ is:

$$\hat{Y}_{dw} = \sum_s w_k\, y_{dk} \quad (5)$$

where $w_k = d_k g_k$.

Thus

$$\hat{Y}_{dw} = \sum_s d_k g_k\, y_{dk} \quad (6)$$

This can be written in regression form as in equation (13).

In section 2 we discuss how to find the design weights w_k for a given sample s given a distance function. The expectation, variance, and variance estimation of the domain calibration estimator as well as the relationship of the domain calibration estimator to the generalized regression (GREG) estimator is also discussed. In section 3, we discuss the approach for finding the best sample design for the domain calibration estimator using appropriate design weights. Section 4 discusses the approach for the derivation of optimum stratum sample sizes that would minimize the variance of the domain calibration estimator and reduce the objective function under six different criteria. In section 5 we present data analysis and discussion. Section 6 presents the conclusions for the paper.

2. Derivation of Calibration Estimators for Domain

Given a sample s, we want to find w_k close to d_k based on a distance function $D(w,d)$ subject to the constraint in equation (4). This is an optimization problem where we wish to minimize

$$\varphi(w_1,\dots,w_n,\lambda) = \sum_s D(w_k,d_k) - \lambda\left(\sum_s w_k x_{dk} - \hat{X}_d\right) \quad (7)$$

using the method of Lagrange Multipliers.

We will derive our calibration weights using the chi-squared distance $(w-d)^2/2qd$ where q is a tuning parameter that can be manipulated to achieve the optimal minimum of equation (7).

However, in practice, it should be noted that the choice of distance function $D(w,d)$ depends on the statistician and the problem considered.

$$\text{Let } D(w_k,d_k) = (w_k - d_k)^2/2q_k d_k \quad (8)$$

Thus equation (7) becomes

$$\varphi = \sum_s (w_k - d_k)^2/2q_k d_k - \lambda\left(\sum_s w_k x_{dk} - \hat{X}_d\right) \quad (9)$$

Differentiating equation (9) with respect to w_k and equating to zero we have

$$w_k = d_k[1 + \lambda q_k x_{dk}^T] \quad (10)$$

substituting equation (10) into (4) and solving for λ we have

$$\lambda = \frac{\sum_s w_k x_{dk} - \sum_s d_k x_{dk}}{\sum_s d_k q_k x_{dk} x_{dk}^T}$$

$$\lambda = (\hat{X}_{d,w} - \hat{X}_d)\alpha^{-1} \quad (11)$$

where $\alpha = \sum_s d_k q_k x_{dk} x_{dk}^T$.

substituting (11) into (10) we have

$$w_k = d_k\left\{1 + q_k x_{dk}^T\left\{(\hat{X}_{d,w} - \hat{X}_d)\alpha^{-1}\right\}\right\} \text{ and so}$$

$$\hat{Y}_{d,w} = \sum_s\left\{d_k\left\{1 + q_k x_{dk}^T(\hat{X}_{d,w} - \hat{X}_d)\alpha^{-1}\right\}y_{dk}\right\}$$

$$\hat{Y}_{d,w} = \sum_s d_k y_{dk} + (\hat{X}_{d,w} - \hat{X}_d)\hat{\beta}_d \quad (12)$$

Where

$$\hat{\beta}_d = \frac{\sum_s d_k q_k x_{dk}^T\, y_{dk}}{\sum_s d_k q_k x_{dk} x_{dk}^T}$$

Following from equation (1): That is $\hat{Y}_{dHT} = \sum_s d_k y_{dk}$

Thus

$$\hat{Y}_{d,w} = \hat{Y}_{dHT} + (\hat{X}_{d,w} - \hat{X}_d)\hat{\beta}_d \quad (13)$$

Equation (13) is our proposed calibration estimator. It is a version of the Generalized Regression Estimator (GREG - estimator). This implies that the GREG-estimator is a special case of the calibration estimator in equation (5). Our result in (13) conforms to the GREG-estimator proposed by [2]. In fact, the GREG-estimator is a special case of the calibration estimator when the chosen distance function is the chi-square distance (see [4]).

2.1. Variance and Variance Estimation

We will follow the procedure proposed by[4] to derive an approximate variance for our proposed calibration estimator $\hat{Y}_{d,w}$.

To find the expectation and variance of $\hat{Y}_{d,w}$, we use the linearization technique to find an approximation of $E_P(\hat{Y}_{d,w})$ and $V_P(\hat{Y}_{d,w})$ with respect to a probability design P. Let β_d be the population level version of $\hat{\beta}_d$. Then a linear approximation of $(\hat{Y}_{d,w})$ is:

$$(\hat{Y}_{d,w}) = \hat{Y}_{dHT} + \beta_d(\hat{X}_{d,w} - \hat{X}_d) + (\hat{\beta}_d - \beta_d)(\hat{X}_{d,w} - \hat{X}_d) \quad (14)$$

Where the first term is of order $O_P(1)$, the second term is

of order $O_P\left(n^{-\frac{1}{2}}\right)$ and the last term is of order $O_P(n^{-1})$ as shown by [4]. Consequently, the last term can be omitted since it is of order $O_P(n^{-1})$. Thus, we can rewrite (14) as:

$$\left(\hat{Y}_{d,w}\right) = \hat{Y}_{dHT} + \beta_d\left(\hat{X}_{d,w} - \hat{X}_d\right) \tag{15}$$

Using equation (15), the design-based expectation of $\hat{Y}_{d,w}$ is:

$$E_P\left(\hat{Y}_{d,w}\right) = E_P\left\{\hat{Y}_{dHT} + \beta_d\left(\hat{X}_{d,w} - \hat{X}_d\right)\right\}$$
$$E_P\left(\hat{Y}_{d,w}\right) = E_P\left(\hat{Y}_{dHT}\right)$$
$$E_P\left(\hat{Y}_{d,w}\right) = E_P\left(\sum_s d_k y_{dk}\right)$$
$$E_P\left(\hat{Y}_{d,w}\right) = \sum_s y_{dk}\, E(d_k) = \sum_s y_{dk} = Y_d \tag{16}$$

Thus $\hat{Y}_{d,w}$ is an approximately design-unbiased estimator of the domain total Y_d. Note that

$$E(d_k) = E\left(\frac{1}{\pi_k}\right) = \sum_{k=1}^{N}\left(\frac{1}{\pi_k}\right) = 1 \text{ since } \sum P(s) = 1$$

Again using equation (15), the design-based asymptotic variance of $\hat{Y}_{d,w}$ is

$$V_P\left(\hat{Y}_{d,w}\right) = V_P\left\{\hat{Y}_{dHT} + \beta_d\left(\hat{X}_{d,w} - \hat{X}_d\right)\right\} \tag{17}$$
$$V_P\left(\hat{Y}_{d,w}\right) = V_P\left(\hat{Y}_{dHT} - \beta_d \hat{X}_d\right)$$
$$V_P\left(\hat{Y}_{d,w}\right) = V_P\left(\sum_S d_k y_{dk} - \beta_d \sum_s d_k x_{dk}\right)$$
$$V_P\left(\hat{Y}_{d,w}\right) = V_P\left[\sum_s d_k\left(y_{dk} - x_{dk}^T \beta_d\right)\right]$$
$$V_P\left(\hat{Y}_{d,w}\right) = \sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl} - \pi_k\pi_l)\left(d_k\left(y_{dk} - x_{dk}^T\beta_d\right)\right) \times \left(d_l\left(y_{dk} - x_{dk}^T\beta_d\right)\right)$$
$$V_P\left(\hat{Y}_{d,w}\right) = \sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl} - \pi_k\pi_l)(d_k E_{dk})(d_l E_{dl})$$

where $E_{dk} = y_{dk} - x_{dk}^T\beta_d$

The variance estimator is;

$$V_P\left(\hat{Y}_{d,w}\right) = \sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl} - \pi_k\pi_l)(d_k E_{dk})(d_l E_{dl})$$
$$V_P\left(\hat{Y}_{d,w}\right) = \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{d_k d_l}{d_{kl}} - 1\right)E_{dk}\ E_{dl} \tag{18}$$

Note that the $V\left\{\beta_d\left(\hat{X}_{dw}\right)\right\} = 0$, a consistent and approximate unbiased estimator of variance (18) is:

$$\hat{V}_d\left(\hat{Y}_{dw}\right) = \sum_{k=1}^{N}\sum_{l=1}^{N} D_{kl}(w_k e_{dk})\,(w_k e_{dl})$$

where $e_{dk} = y_{dk} - x_{dk}^T\hat{\beta}_d$.

It should be noted that, it is acceptable to use the design weights d_k in the variance estimation as in equation (18), but[4] suggest that the calibration weights w_k be used in equation (18) as this makes the variance estimator both design-consistent and nearly model unbiased. Moreover, since the calibration estimator is asymptotically equivalent to the GREG-estimator, it can be inferred that calibration estimators are more efficient compared to HT-estimator if there is a strong correlation between y and x [2].

3. Sample Design for the Calibration Estimator

Consider a stratified random sampling design with H strata and such that n_h elements are considered from N_h in stratum $h, h = 1,2,\ldots,H$. Then, the design weights needed for the point estimation are $d_k = N_h/n_h$ for all k in stratum $h, k = 1,2,\ldots,N_h$. However, the design weights d_{kl} needed for the variance estimation if $k \neq l$ and both k and l are in different strata, say stratum h and stratum h' is:

$$d_{kl} = \frac{N_h}{n_h}\frac{N_{h'}}{n_{h'}} \tag{19}$$

Using equation (18): $\sum_{h=1}^{H}\sum_{k=1}^{N_h}(d_k d_l/d_{kl} - 1)\,E_k E_l$

Then, we have;

$$\sum_{h=1}^{H}\sum_{h=1}^{N_h}\left\{\left[\left(\frac{N_h}{n_h}\right)^2\frac{N_{h'}}{n_{h'}} - \frac{N_h}{n_h}\frac{N_{h'}}{n_{h'}}\right]\Big/\frac{N_h}{n_h}\frac{N_{h'}}{n_{h'}}\right\}E_k E_l$$
$$\sum_{h=1}^{H} N_h^2\frac{(1-f_h)}{n_h}E_k E_l \tag{20}$$

It should be noted that in calibration, it is assumed that elements respond independently so that $d_{kl} = d_k.d_l$. Consequently, from the following Theorem in stratified sampling design according to [3];

Theorem

If the samples are drawn independently in different strata,

$$V(\bar{y}_{st}) = \sum_{h=1}^{H} w_h^2 V(\bar{y}_h)$$

where $V(\bar{y}_h)$ is the variance of $\bar{y}_h$ over repeated samples from stratum h.

Since

$$\bar{y}_{st} = \sum_{h=1}^{H} w_h\bar{y}_h$$

$\bar{y}_{st}$ is a linear function of the $\bar{y}_h$ with fixed weights w_h. Hence we may quote the result in statistics for the variance of a linear function

$$V(\bar{y}_{st}) = \sum_{h=1}^{H} w_h^2 V(\bar{y}_h) + 2\sum_{h=1}^{H}\sum_{j>h}^{H} w_h w_j\, cov\left(\bar{y}_h, \bar{y}_j\right)$$

But since samples are drawn independently in different strata, all covariance terms varnish [3 p. 92].

We have variance estimator of (18) as:

$$\hat{V}_P\left(\hat{Y}_{d,w}\right) = \sum_{h=1}^{H} N_h^2\frac{(1-f_h)}{n_h}\,(e_k e_l)$$

$$\hat{V}_P(\hat{Y}_{d,w}) = \sum_{h=1}^{H} N_h^2 \frac{(1-f_h)}{n_h} S_h^2 + 2\sum_{h=1}^{H}\sum_{l>h}^{H} N_h N_l cov(e_k e_l)$$

$$\hat{V}_P(\hat{Y}_{d,w}) = \sum_{h=1}^{H} N_h^2 \frac{(1-f_h)}{n_h} S_h^2 \quad (21)$$

4. Optimal Sample Allocations

We shall now deduce the optimum n ($n_{h,opt}$), that minimize the variances of the proposed calibration estimators for a specified cost, or that minimize the cost for a specified variance.

Let us consider the simple linear sampling cost function of the form:

$$C = c_0 + \sum_{h=1}^{H} c_h n_h \quad (22)$$

where c_0 is the overhead cost and c_h is the cost per unit of obtaining the necessary information in h-th stratum. In this paper, we shall consider the following allocation methods: optimum allocation, Neyman allocation, optimal power allocation, Neyman power allocation, square root allocation and Neyman square root allocation.

(i) Optimum allocation

The problem of optimum allocation consists in minimizing the sampling variance for a given overall sampling cost of the survey or minimizing the overall sampling cost for specified sampling variance.

Let us consider the simple linear sampling cost function of equation (22), the corresponding Lagragian is:

$$G = \sum_{h=1}^{H} N_h^2 (1-f_h)\frac{S_h^2}{n_h} + \lambda\{\sum_{h=1}^{H} c_h n_h + c_0 - C\} \quad (23)$$

Differentiating (23) with respect to n_h and λ and equating to zero we have respectively

$$\lambda c_h n_h^2 = N_h^2 S_h^2$$

and

$$C - c_0 = \sum_{h=1}^{H} c_h n_h$$

Thus

$$n_h = \frac{N_h S_h}{\sqrt{\lambda c_h}} \quad (24)$$

and

$$\sqrt{\lambda} = \sum_{h=1}^{H} c_h N_h S_h \Big/ \sqrt{c_h}\,(C - c_0)$$

Finally to obtain a solution for n_h, we substitute for $\sqrt{\lambda}$ into (24) as follows:

$$n_{h,opt} = \frac{(C-c_0)N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{H} c_h N_h S_h/\sqrt{c_h}} \quad (25)$$

(ii) Neyman allocation

If the cost per unit is the same in all strata, (that is, $c_h = c, h = 1,2,\ldots,H$), then (25) reduces to

$$n_{h,opt} = \frac{(C-c_0)N_h S_h}{c\sum_{h=1}^{H} N_h S_h} \quad (26)$$

The type of allocation of (26) where $c_h = c$ (that is, where the cost per unit is the same in all strata) is called the Neyman allocation.

(iii) Optimal power allocation

The power allocation was first considered by [1]. He considered a compromise allocation between equal allocation and Neyman allocation in which the within stratum sample size is proportional to N_h^p and called it power allocation. Suppose that a stratified random sampling is to be selected. Let N_h be some measure of size or importance for the h th stratum. It is desired to determine stratum sample sizes n_h, that the loss function

$$L = \sum_{h=1}^{H}\{N_h^p CV(\hat{Y}_H)\}^2 \quad (27)$$

is minimized subject to the constraint $C = c_0 + \sum_{h=1}^{H} c_h n_h$ where $CV^2(\hat{Y}_h) = V_p(\hat{Y}_h)/\hat{Y}_h^2$ and p is a constant in the range $0 \le p \le 1$ and is called the power of the allocation.

Following from our sample design, the loss function is

$$L = \sum_{h=1}^{H}\{N_h^p CV(\hat{Y}_{d,w})\}^2$$

$$L = \sum_{h=1}^{H}\left\{\sum_{h=1}^{H} N_h^2 S_h^2/n_h - \sum_{h=1}^{H} N_h S_h^2\right\}\frac{(N_h^p)^2}{\hat{Y}_h^2}$$

The corresponding lagragian is

$$G = \sum_{h=1}^{H}\left\{\sum_{h=1}^{H} N_h^2 S_h^2/n_h - \sum_{h=1}^{H} N_h S_h^2\right\}\frac{(N_h^p)^2}{\hat{Y}_h^2} + \lambda\{\sum_{h=1}^{H} c_h n_h + c_0 - C\} \quad (28)$$

Differentiating (28) with respect to n_h and λ and equating to zero we have

$$n_h^2(\lambda c_h \hat{Y}_h^2) = N_h^2 S_h^2 (N_h^p)^2$$

and

$$C - c_0 = \sum_{h=1}^{H} c_h n_h$$

Thus,

$$n_h = \frac{N_h S_h N_h^p}{\hat{Y}_h\sqrt{\lambda c_h}} \quad (29)$$

and

$$\sqrt{\lambda} = \frac{\sum_{h=1}^{H} c_h N_h S_h N_h^p}{(C-c_0)\hat{Y}_h\sqrt{c_h}}$$

Finally to obtain a solution for n_h, we substitute for $\sqrt{\lambda}$ into (29) as follows:

$$n_{h,opt} = \frac{(C-c_0)S_h N_h N_h^p/\sqrt{c_h}}{\sum_{h=1}^{H} c_h S_h N_h N_h^p/\sqrt{c_h}} \quad (30)$$

The exponent p is called the power of the allocation. The choice of p results in significantly different allocations, for example, if the cost per unit is the same across strata (that is,

$c_h = c, h = 1,2, \ldots, H$) and setting $p = 1$, we obtain the Neyman allocation of (26). Choosing a value of p between 0 and 1 can be viewed as a compromise allocation between Neyman allocation and the equal allocation.

(iv) Neyman power allocation

If the cost per unit is the same across strata, then;

$$n_{h,opt} = \frac{(C-c_0)S_h N_h N_h^p}{c\sum_{h=1}^{H} S_h N_h N_h^p} \quad (31)$$

(v) Square root allocation

The square root allocation is a special case of the power allocation. When the power of the allocation p is set to one-half (that is, setting $p = 0.5$), we obtain

$$n_{h,opt} = \frac{(C-c_0)N_h S_h \sqrt{N_h}/\sqrt{c_h}}{\sum_{h=1}^{H} c_h N_h S_h \sqrt{N_h}/\sqrt{c_h}} \quad (32)$$

(vi) Neyman square root allocation

Again, if the cost per unit is the same across strata (that is, $c_h = c, h = 1,2, \ldots, H$), and the power of the allocation p is set to one-half (that is, setting $p = 0.5$), then, we obtain what may be called the Neyman square root allocation as:

$$n_{h,opt} = (C - c_0)N_h S_h \sqrt{N_h}/c\sum_{h=1}^{H} N_h S_h \sqrt{N_h} \quad (33)$$

5. Data Analysis and Discussion

5.1. Background and Analytical Set-up

The data used is obtained from the 2005 socio-economic household survey of Akwa Ibom State conducted by the ministry of economic development, Uyo, Akwa Ibom State, Nigeria.

The study variable, y, represents the household expenditure on food and auxiliary variable, x, represents the household income. The statistic of interest is the total cost of food for household and its corresponding estimator for male and female heads of household.

The population of household heads was stratified into two strata that constitute the domains; as the male household heads and the female household heads respectively. For the population of individual household heads, we want a separate estimates for male and female household heads defined as two domains of the population. The number of the male household heads and female household heads in the survey are known. We used the calibration estimator for the domain total $\hat{Y}_{d,w}, d = 1,2$ and the following formulation is specified: The number of male household heads, N_1 and female household heads, N_2 are known and the auxiliary vector has two possible values; namely, $\boldsymbol{x}_k = (1,0)^T$ for all male household heads and $\boldsymbol{x}_k = (0,1)^T$ for all female household heads. The population total of the auxiliary vector $\boldsymbol{x}_k$ is $(N_1, N_2)^T$ which is also known and $q_k = 1$ for all k.

An assisting model of the form $y_h = \beta_0 + \beta_1 x_h + e_h$ was designed for the calibration estimators, where h is the number of strata (domains) and e_h are independently generated by the standard normal distribution.

5.2. The Sampling Design Variance Estimation

To obtain an optimum value of n_h that minimizes the design variance $V_P(\hat{Y}_{d,w})$, a population was generated with the following parameters: $C = 500, c_0 = 100, c = 0.4, c_1 = 0.5, c_2 = 0.3, S_1^2 = 0.3262,\ S_1 = 0.5711, \rho = 0.7670, N_1 = 7{,}396;\ N_2 = 1{,}553; N = 8{,}949; S_2^2 = 0.4326, S_2 = 0.6577$. Table 1 shows the summary of values of n_h for the six allocation criteria. The variance for the domain calibration estimator using the optimum values of n_h from the six different allocation criteria are presented in table 2.

Table 1. Optimum Value of n_h

Stratum	OA	NA	OPA	NPA	SRA	NSRA
1	674	805	770	952	737	900
2	210	195	50	48	105	100
Total	884	1,000	820	1,000	842	1,000

Table 2. Optimum Variance

Allocation Method	Stratum 1	Stratum 2	Total
Optimum Allocation	24,061.3212	4,296.4987	28,357.8199
Neyman Allocation	19,753.1468	4,678.6777	24,431.8245
Optimum Power Allocation	20,760.6796	20,195.143	40,955.8233
Neyman Power Allocation	16,33.4985	21,064.601	37,395.0993
Square Root Allocation	21,798.2880	9,264.8253	31,063.1133
Neyman Square Root Allocation	17,413.4317	9,761.6579	27,175.0896

The variance estimator from the stratified random sampling design is;

$$V_P(\hat{Y}_{d,w}) = \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_h^2}{n_h}$$

where $h = 1,2$ and $\rho_{xy} = 0.7670$ and S_h^2 is the stratum variance of the residuals e_{dk} where $e_{dk} = y_{dk} - \boldsymbol{x}_k^T\hat{\beta}_d$. The optimum value of n_h for the Neyman allocation gave the minimum variance. The results of the design variance estimation are presented in table 3.

Table 3. Variance Estimation

Stratum	N_h	n_h	$N_h - n_h$	S_h^2	$N_h(N_h - n_h)\frac{S_h^2}{n_h}$
1.	7,396	805	6,591	0.3262	19,753.1468
2.	1,553	195	1,358	0.4326	4,678.6777
Total	8,949				24,431.8245

5.3. Comparisons with Greg-estimator

To compare the performance of each estimator we use the following criteria; bias (B), relative bias (RB), mean square error (MSE), average length of confidence interval (AL) and the coverage probability (CP) of $\hat{Y}_{d,w}$. Let $\hat{Y}_{d,w}^{(m)}$ be the estimate of $\hat{Y}_{d,w}$ in the m -th simulation run; $m = 1,2, \ldots, M(= 2{,}500)$ we define

i. $B(\hat{Y}_{d,w}) = \bar{\hat{Y}}_{d,w} - \hat{Y}_{d,w}^{(m)}$

where $\bar{\hat{Y}}_{d,w} = \frac{1}{M_d}\sum_{m=1}^{M_d} \hat{Y}_{d,w}^{(m)}$

ii. $RB(\hat{Y}_{d,w}) = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{Y}_{d,w}^{(m)} - \bar{\hat{Y}}_{d,w}}{\bar{\hat{Y}}_{d,w}}\right)$

iii. $MSE(\hat{Y}_{d,w}) = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{Y}_{d,w}^{(m)} - \bar{\hat{Y}}_{d,w}\right)^2$

iv. $AL(\hat{Y}_{d,w}) = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{Y}_{U,d,w}^{(m)} - \hat{Y}_{L,d,w}^{(m)}\right)$

where $\hat{Y}_{U,d,w}^{(m)}$ and $\hat{Y}_{L,d,w}^{(m)}$ are the upper and lower confidence limit of the corresponding confidence interval.

v. $AL(\hat{Y}_{d,w}) = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{Y}_{L,d,w}^{(m)} < \hat{Y}_{d,w} < \hat{Y}_{U,d,w}^{(m)}\right)$

Coverage probability of 95% confidence interval is the ratio of the number of times the true domain total is included in the interval to the total number of runs or the number of replicates.

For each estimator of $\hat{Y}_{d,w}$, a 95% confidence interval $(\hat{Y}_{L,d,w}, \hat{Y}_{U,d,w})$ is constructed, where

$$\hat{Y}_{L,d,w} = \hat{Y}_{d,w}^{(m)} - I.96\sqrt{V(\hat{Y}_{d,w}^{(m)})}$$

and $$\hat{Y}_{U,d,w} = \hat{Y}_{d,w}^{(m)} + I.96\sqrt{V(\hat{Y}_{d,w}^{(m)})}$$

where $\hat{Y}_{L,d,w}$ is the lower confidence limit, $\hat{Y}_{U,d,w}$ is the upper confidence limit and $V(\hat{Y}_{d,w}^{(m)}) = \frac{1}{M_d - 1}\sum_{m=1}^{M_d}\left(\hat{Y}_{d,w}^{(m)} - Yd,w2\right.$.

The analytical study was conducted using the R-statistical package. There were $M = 2{,}500$ runs in total. For the m-th run $(m = 1,2,\ldots,M)$, a Bernoulli sample is drawn where each unit is selected into the sample independently, with inclusion probability $\pi_k = N_h/n_h$ where $h = 1,2$. Following the results of analysis for optimum stratum sample sizes, we fixed $n_1 = 805$ and $n_2 = 195$ and the corresponding calibration estimators of the domain totals were computed. For simplicity, the tuning parameter q_k was set to unity $(q_k = 1)$.

For each estimator of $\hat{Y}_{d,w}$, a 95% confidence interval $(\hat{Y}_{L,d,w}, \hat{Y}_{U,d,w})$ is constructed, where $\hat{Y}_{L,d,w}$ is the lower confidence limit, and $\hat{Y}_{U,d,w}$ is the upper confidence limit.The results of the analysis are given in table 4.

Table 4. Comparison of estimators from analytical study

Estimator	B	RB	MSE	AL	CP
$\hat{Y}_{d,GREG}$	0.0096	0.0632	5896	1283.50	0.982
$\hat{Y}_{d,w}$	0.0082	0.0162	2632	876.32	0.867

5.4. Discussion

An assisting model of the form $y_h = \beta_0 + \beta_1 x_h + e_h$ was designed where h is the number of strata (domains) and $e_h \sim N(0, \sigma_{e_h}^2)$. The results of the residual diagnostics showed the R^2 value as 0.588 indicating that the model is significant and that the calibration estimators are unbiased with respect to the sampling design. The correlation between the study variable y and the auxiliary variable x is $\rho_{xy} = 0.7670$ is strong and sufficient implying that the calibration estimators would provide better estimates of the domain totals.

The Neyman allocation criterion provides the optimum stratum sample sizes $n_{1,opt} = 805$ and $n_{2,opt} = 195$ that minimized the variance of the calibration estimators as reflected in table 2.

The design strata estimates are 19,753.1468 and 4,678.6777 for stratum 1 and stratum 2 respectively. Similarly, the variance estimate is 24,431.8245. Following from the above estimates, we deduced that the design strata estimates sum up to the finite population estimates.

Analysis for the comparison of performance of estimators showed that the biases of 0.82 percent and 0.96 percent respectively for the calibration estimator and the GREG-estimator are negligible. But the bias of the GREG-estimator though negligible is the most biased among the estimators considered.

The relative bias for the calibration estimator is relatively smaller than that of the GREG-estimator. The variance for the GREG-estimator is significantly larger than the variance of the calibration estimators, as is indicated by their respective mean square errors in table 4. The average length of the confidence interval for the calibration estimator is significantly smaller than that of the GREG-estimator. The coverage probability of the calibration estimator is also smaller than that of the GREG-estimator. These results showed that there is greater variation in the estimates made by the GREG-estimator than the calibration estimator.

In general, the domain calibration estimator is more efficient than the GREG-estimator and the variance reduction is about 50 percent which is consistent with theory as is reflected by the high population correlation between the study variable y and the auxiliary variable x.

6. Conclusions

Nonresponse as an aspect in almost every type of sample survey creates problems for estimation which cannot simply be eliminated by increasing sample size. This phenomenon of nonresponse in a sample survey reduces the precision of parameters estimates and increases bias in estimates resulting in larger mean square error, thus ultimately reducing their efficiency.

Sample surveys have long been used as cost-effective means for data collection. Such data is used to provide suitable statistics not only for the population targeted by the survey but also for a variety of subpopulations called domains of study. Sample designs and in particular sample sizes are chosen so as to provide reliable estimates for domains of study.

One of the main objectives of sample survey is the computation of estimates of means and totals for specific domains of interest. The reliability of the associated estimates depends on the variability of the sample size as well as on the study variables, y of interest. But budget and other constraints usually prevent the allocation of

sufficiently large samples to domains to provide reliable estimates using traditional statistical techniques. This problem of optimal allocation of sample sizes for domain estimation has received less attention than merited in the statistical sample survey theory literature.

This paper addressed these problems by proposing calibration estimators for totals of domains of study and developed an approach for finding the best sample design for the domain calibration estimators subject to a cost constraint in the context of stratified random sampling design (STRS) where domains constitute strata in the sampling design to obtain optimal stratum sample sizes that minimized the variances of the proposed domain calibration estimators and reduced the objective function.

The efficacy of our proposed calibration estimator was tested through a real data analysis. Five performance criteria, namely; bias (B), relative bias (RB), mean square error (MSE), average length of confidence interval (AL) and coverage probability (CP) were used to compare the relative performances of our proposed domain calibration estimator against the traditional GREG-estimator. Results of the analytical study using real data showed that our proposed calibration estimator is substantially superior to the traditional GREG-estimator with relatively small bias, mean square error and average length of confidence interval.

REFERENCES

[1] M. D. Bankier, *Power allocation: determining sample sizes for subnational areas*. The American Statistician 12 (1988), pp. 174-177.

[2] C. M. Cassell, C. E. Sarndal and J. H. Wretman, *Some results on generalized difference estimation and generalized regression estimation for finite populations.* Biometrika 63 (1976) pp. 615-620.

[3] W.G. Cochran, *Sampling Techniques*, Wiley and Sons, New York, 1977.

[4] J. C. Deville and C. E. Sarndal, *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association, 87 (1992), pp. 376-382.

[5] V.M. Estavao and C.E. Sarndal, *Survey estimates by calibration on complex auxiliary information*, International Statistical Review, 74 (2006), pp. 127-147.

[6] P. J. Farrell and S. Singh, *Model-assisted higher order calibration of estimators of variance,* Aust. and New Zealand Journal of Statistics, 47 (2005), pp. 375-383.

[7] W. Gamrot, *Estimation of a domain total under nonresponse using double sampling.* Statistics in Transition, 7 (2006), pp. 831-840.

[8] D. G. Horvitz and D. J. Thompson, *A generalization of sampling without replacement from a finite universe.* Journal of the American Statistical Association, 47 (1952), pp. 663-687.

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

[10] G. E. Montanari and M. G. Ranalli, *Nonparametric model calibration estimation in survey sampling,* Journal of the American Statistical Association, 100 (2005), pp. 1429-1442.

[11] S. Singh, *Survey statistician celebrate golden jubilee year-2003 of the linear regression estimato*r, Metrika 2006 pp.1-18.

[12] Singh, *Calibrated empirical likelihood estimation using a displacement function: Sir R. A. Fisher's Honest Balance*. INTERFACE Pasadena, CA, USA. 2006.

[13] C. Wu and R. R. Sitter, *A Model-calibration approach to using complete auxiliary information from survey data,* Journal of the American Statistical Association, 96 (2001), pp. 185-193.