*Aligarh Journal of Statistics*
*Vol. 35(2015,* 117-132

# ESTIMATION FOR DOMAINS IN STRATIFIED SAMPLING DESIGN IN THE PRESENCE OF NONRESPONSE

E. P. Clement, G. A. Udofia and Ekaette I. Enang

## ABSTRACT

An analytical approach for finding the best sampling design subject to a cost constraint is developed. We consider stratified random sampling design when elements of the inclusion probabilities are not equal but are in same stratum and proposed estimators of totals for domains of study under non-response in the context of calibration estimation. We derived optimum stratum sample sizes for a given set of unit costs for the sample design and compared empirically the relative performances of the proposed calibration estimators with a corresponding global estimator. Analysis and evaluation are presented.

## 1. INTRODUCTION

In sample survey, separate estimates of a parameter may be required for subpopulations into which a population is divided without separately sampling from these subpopulations. Such subpopulations are called domains of study (Hidiroglou and Patak 2006). The method of estimating the domain parameters is called domain estimation.

Yates (1953) first considered in detail some of the problems associated with the estimation of domain totals, means and proportions in the case of a single-stage simple random sampling. He noted that the variance of an estimator of a domain parameter is increased by the fact that the number of the domain elements, and hence the number of those elements that can fall in a random sample of a fixed size, is unknown before the start of the survey. Hartley (1959) gave a derivation of Yates' results in multi-stage sampling. Hartley's paper (1959) is one of the first attempts to unify the theory of domain estimation. Hartley provided the theory for a number of sample designs where domain estimation was of interest. His paper mostly discussed estimations that did not make use of auxiliary information. He did, however, consider the case of ratio estimation where population totals were known for the domains.

Udofia (2002) extended Yates' results to double sampling for probability proportional to size (PPS) when information on the size, X, of each sampling unit is unknown.

Torabi, Datta and Rao (2009) proposed an empirical Bayes estimation of domain means under nested error linear regression model with measurement errors in the covariates.

The problem of allocation of resources when domains of study are of primary interest is discussed by Cochran (1977).

However, despite these vast extensions of Yates results, the phenomenon of non-response and its problems in domain estimation have not been addressed.

In many human surveys, information is in most cases not obtained from all the units in the survey even after some call-backs. An estimate obtained from such incomplete data may be misleading especially when the respondents differ from the non-respondents because the estimate can be biased.

Non-response always exists when surveying human populations as people hesitate to respond in surveys; and increases notably while studying sensitive issues like family size as in a case of survey of fishing communities in Umon Island, Nigeria undertaken in 1993. Non-response as an aspect in almost every type of sample survey creates problems for estimation which cannot simply be eliminated by increasing sample size.

The phenomenon of non-response 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 non-response was first introduced by Deville and Sarndal (1992) for the estimation of finite population characteristics like means, ratios and totals. This calibration approach requires the formulation of suitable auxiliary variables. The calibration approach provides a unified treatment of the use of auxiliary information in surveys with non-response. In the presence of powerful auxiliary information, the calibration approach meets the objectives of reducing both the sampling error and the non-response error. This article is an attempt to extend Yates' results to stratified sampling design for domain estimation in the presence of non-response.

## 2. SAMPLE DESIGN AND ESTIMATION

### 2.1 Domain Characteristics

Consider the finite population under study $U$ of size $N$ divided into $D$ domains; $U_1, U_2, ..., U_D$ of sizes $N_1, N_2, ..., N_D$ respectively. Domain membership of any population unit is unknown before sampling. It is assumed that domains are quite large. Following from Gamrot (2006), for a typical $d^{th}$ domain $U_d$ several characteristics may be defined including the domain total:

$$Y_{U_d} = \sum_{U_d} y_{dk} \tag{1}$$

Domain mean

$$\bar{Y}_{U_d} = \frac{1}{N_d} \sum_{U_d} y_{dk} \tag{2}$$

Domain variance

$$S^2_{U_d}(Y) = \frac{1}{N_d - 1} \sum_{k \in U_d} (y_{dk} - \bar{Y}_{U_d})^2 \tag{3}$$

And domain covariance between two characteristics *X* and *Y*

$$C_{U_d}(X, Y) = \frac{1}{N_d - 1} \sum_{k \in U_d} (x_{dk} - \bar{X}_{U_d})(y_{dk} - \bar{Y}_{U_d}) \tag{4}$$

In this article the estimation of domain totals is considered.

### 2.2 Domain Estimation by Calibration

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 unites labels 1,2,...,N. Let $y_k, k = 1,2,...,N$ be the study variable and $x_k, k = 1,2,...,N$ 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 $s = \{1,2,...,n\} \in U_d$ 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) estimator

$$\hat{Y}_{dHT} = \sum_{s} d_k y_{dk} \tag{5}$$

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

An attractive property of the HT-estimator is that it is guaranteed to be unbiased regardless of the sampling design *P* (Horvitz and Thompson 1952). It variance under *P* is given as:

$$V_P(\hat{Y}_{HT}) = \sum_{k=1}^{N} \sum_{l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \tag{6}$$

Suppose there are $x_k, k = 1,2,...,N$ auxiliary variables at unit $k$ and $x_k = x_1,...,x_n,...,x_N$ may or may not be known a priori. $X_d = \sum_{s} x_{dk}$ is the domain total for ***X***, and is known a priori. Ideally, we would like

$$\hat{X}_d = \sum_{s} d_k x_{dk} \tag{7}$$

but often times this is not true.

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

$$\hat{X}_{d,w} = \sum_{s} w_k x_{dk} = \sum_{U_d} x_{dk} \tag{8}$$

Expression (8) 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:

$$\hat{Y}_{d,w} = \sum_{S} w_k y_{dk} \tag{9}$$

Where $w_k = d_k g_k$.

Thus $$\hat{Y}_{dw} = \sum_{S} d_k g_k y_{dk} \tag{10}$$

This can be written in regression form as:

$$\hat{Y}_{d,w} = \hat{Y}_{dHT} + (\hat{X}_{d,w} - \hat{X}_d)\hat{\beta}_d \tag{11}$$

where $\hat{\beta}_d = \dfrac{\sum_S d_k q_k x_{dk}^T y_{dk}}{\sum_S d_k q_k x_{dk} x_{dk}^T}$

And its variance estimator is;

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

$$V_P(\hat{Y}_{d,w}) = \sum_{k=1}^{N}\sum_{l=1}^{N}(\frac{d_k d_l}{d_{kl}} - 1)E_{dk}E_{dl} \qquad (12)$$

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

### 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 $h$, $h = 1,2,...,H$. Then, the design weights needed for the point estimation are $d_k = \frac{N_h}{n_h}$ for all $k$ in stratum $h, k = 1,2,...,N_h$. However, the design weights $d_{kl}$ needed for the variance estimation if $k \neq l$ and both $k$ and $l$ are in stratum $h$ is: $d_{kl} = \frac{N_h}{n_h}\left(\frac{N_h - 1}{n_h - 1}\right)$

Using equation (12): $\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(\frac{dkdl}{dkl} - 1\right)E_k E_l$

Then we have;

$$V_P(\hat{Y}_{d,w}) = \sum_{h=1}^{H}\sum_{k=1}^{N_h}\left\{\left(\frac{N_h}{n_h}\right)^2\left(\frac{N_h - 1}{n_h - 1}\right) - \frac{\frac{N_h}{n_h}\left(\frac{N_h - 1}{n_h - 1}\right)}{\frac{N_h}{n_h}\left(\frac{N_h - 1}{n_n - 1}\right)}\right\}E_k E_l$$

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

$$V_P(\hat{Y}_{d,w}) = \sum_{h=1}^{H}\sum_{k=1}^{N_h}\frac{N_h(N_h - 1)}{n_h}\left[\frac{(N_h - n_h)}{n_h(n_h - 1)}\right]\frac{n_h}{N_h}\left(\frac{n_h - 1}{N_h - 1}\right)E_k E_l$$

$$V_P(\hat{Y}_{d,w}) = \sum_{h=1}^{H}\sum_{k=1}^{N_h}\frac{N_h}{n_h}\left[\frac{(N_h - n_h)}{N_h}\right]E_k E_l$$

$$V_P(\hat{Y}_{d,w}) = \sum_{h=1}^{H} N_h^2 \frac{(1-f_h)}{n_h} E_k E_l \qquad (13)$$

Therefore our variance estimator of (12) becomes

$$V_P(\hat{Y}_{d,w}) = \sum_{h=1}^{H} N_h^2 \frac{(1-f_h)}{n_h}\operatorname{cov}(e_k e_l) \qquad (14)$$

But $\operatorname{cov}(e_k e_l) = \sigma_h^2 \rho$ and from the principle of SRS $\sigma^2 = \left(\frac{N-1}{N}\right)S^2$.

Therefore, $\sigma_h^2 = \left(\frac{N_h - 1}{N_h}\right)S_h^2$ (15)

and $\operatorname{cov}(e_k e_l) = \left(\frac{N_h - 1}{N_h}\right)S_h^2 \rho$ (16)

Substituting (16) into (14) we have

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

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

$$V_P(\hat{Y}_{d,w}) = \frac{1}{n_h}\left[\sum_{h=1}^{H} N_h^2 S_h^2 \rho - \sum_{h=1}^{H} N_h S_h^2 \rho\right] - \sum_{h=1}^{H} N_h \left(\frac{N_h - 1}{N_h}\right)S_h^2 \rho \qquad (17)$$

### 2.4 Optimal Sample Allocation

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:

$$C = c_0 + \sum_{h=1}^{H} c_h n_h \qquad (18)$$

where $c_0$ is the overhead cost and $c_h$ is the cost per unit of obtaining the necessary information in *h*-th stratum. We shall consider the following allocation methods in this article, namely:

**(i)** Optimum allocation

Using the cost function of (18), $C = c_0 + \sum_{h=1}^{H} c_h n_h$, we have corresponding lagrangian as follows:

$$G_2 = \frac{1}{n_h}\left[\sum_{h=1}^{H} N_h^2 S_h^2 \rho - \sum_{h=1}^{H} N_h S_h^2 \rho\right] - \sum_{h=1}^{H} N_h \left(\frac{N_h - 1}{N_h}\right) S_h^2 \rho + \lambda \left\{\sum_{h=1}^{H} c_h n_h + c_0 - C\right\} \qquad (19)$$

The partial derivatives of (19) with respect to $n_h$ and $\lambda$ are respectively:

$$\frac{\partial G_2}{\partial n_h} = -\frac{[N_h^2 S_h^2 \rho - N_h S_h^2 \rho]}{n_h^2} + \lambda c_h$$

$$\lambda c_h n_h^2 = N_h S_h^2 \rho (N_h - 1)$$

$$n_h = \frac{\sqrt{N_h S_h^2 \rho (N_h - 1)}}{\sqrt{\lambda c_h}} \qquad (20)$$

$$\frac{\partial G_2}{\partial \lambda} = \sum_{h=1}^{H} c_h n_h + c_0 - C$$

$$C - c_0 = \sum_{h=1}^{H} c_h n_h \qquad (21)$$

substituting (20) into (21) and solving for $\lambda$, we obtain

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

Finally to obtain a solution for $n_h$, we substitute for $\sqrt{\lambda}$ into (20) as follows:

$$n_{h,opt} = \frac{(C - c_0) S_h \sqrt{N_h (N_h - 1)} / \sqrt{c_h}}{\sum_{h=1}^{H} c_h S_h \sqrt{N_h (N_h - 1)} / \sqrt{c_h}} \qquad (22)$$

**(ii)** Neyman allocation

If the cost per unit is the same across strata (that is, $c_h = c,\ h = 1,2,...,H$ ) then;

$$n_{h,opt} = \frac{(C - c_0) S_h \sqrt{N_h (N_h - 1)}}{c \sum_{h=1}^{H} S_h \sqrt{N_h (N_h - 1)}} \tag{23}$$

**(iii)** Optimal power allocation

Let the loss function according to Bankier (1988) be

$$L_2 = \sum_{h=1}^{H} \left\{ \frac{1}{n_h} \left( \sum_{h=1}^{H} N_h^2 S_h^2 \rho - \sum_{h=1}^{H} N_h S_h^2 \rho \right) - \sum_{h=1}^{H} N_h S_h^2 \rho \left( \frac{N_h - 1}{N_h} \right) \right\} \left( \frac{N_h^p}{\hat{Y}_h} \right)^2$$

and the corresponding Lagrangian is

$$G_L = \sum_{h=1}^{H} \left\{ \frac{1}{n_h} \left( \sum_{h=1}^{H} N_h^2 S_h^2 \rho - \sum_{h=1}^{H} N_h S_h^2 \rho \right) - \sum_{h=1}^{H} N_h S_h^2 \rho \left( \frac{N_h - 1}{N_h} \right) \right\} \left( \frac{N_h^p}{\hat{Y}_h} \right)^2 + \lambda \left\{ \sum_{h=1}^{H} c_h n_h + c_0 - C \right\} \tag{24}$$

The partial derivatives of (24) with respect to $n_h$ and $\lambda$ are respectively:

$$\frac{\partial G_L}{\partial n_h} = -\frac{[N_h^2 S_h^2 \rho - N_h S_h^2 \rho]}{n_h^2} \left( \frac{N_h^p}{\hat{Y}_h} \right)^2 + \lambda c_h$$

$$\lambda c_h n_h^2 \hat{Y}_h^2 = N_h S_h^2 \rho (N_h - 1)(N_h^p)^2$$

$$n_h = \frac{S_h N_h^p \sqrt{N_h (N_h - 1)\rho}}{\hat{Y}_h \sqrt{\lambda c_h}} \tag{25}$$

$$\frac{\partial G_L}{\partial \lambda} = \sum_{h=1}^{H} c_h n_h + c_0 - C$$

$$C - c_0 = \sum_{h=1}^{H} c_h n_h \tag{26}$$

substituting (25) into (26) and solving for $\lambda$ we obtain

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

Finally to obtain a solution for $n_h$, we substitute for $\sqrt{\lambda}$ into (25) to obtain:

$$n_{h,opt} = \frac{(C-c_0)S_h N_h^p \sqrt{N_h(N_h-1)}/\sqrt{c_h}}{\sum_{h=1}^{H} c_h S_h N_h^p \sqrt{N_h(N_h-1)}/\sqrt{c_h}} \qquad (27)$$

**(iv)** Neyman power allocation

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

$$n_{h,opt} = \frac{(C-c_0)S_h N_h^p \sqrt{N_h(N_h-1)}}{c\sum_{h=1}^{H} S_h N_h^p \sqrt{N_h(N_h-1)}} \qquad (28)$$

**(v)** Square root allocation

If the value of the power of the allocation is set to one-half (*i.e.* 0.5) then

$$n_{h,opt} = \frac{(C-c_0)S_h N_h \sqrt{(N_h-1)}/\sqrt{c_h}}{\sum_{h=1}^{H} c_h S_h N_h \sqrt{(N_h-1)}/\sqrt{c_h}} \qquad (29)$$

**(vi)** Neyman square root allocation

If the cost per unit is the same across strata, and the value of the power of allocation is set to one-half, then, we obtain

$$n_{h,opt} = \frac{(C-c_0)S_h N_h \sqrt{(N_h-1)}}{c\sum_{h=1}^{H} S_h N_h \sqrt{(N_h-1)}} \qquad (30)$$

## 3. DATA ANALYSIS AND DISCUSSION

### 3.1 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, $x_k = (1,0)^T$ for all male household heads and $x_k = (1,0)^T$ for all female household heads. The population total of the auxiliary vector $x_k$ 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.

### 3.2 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 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 | 18,452.5381 | 3,293.2926 | 21,745.8307 |
| Neyman Allocation | 15,148.6151 | 3,586.2351 | 18,734.8502 |
| Optimum Power Allocation | 15,921.2883 | 15,479.701 | 31,400.9895 |
| Neyman Power Allocation | 12,523.7988 | 16,146.145 | 28,669.9442 |

| | | | |
|---|---|---|---|
| Square Root Allocation | 16,717.0263 | 7,101.5452 | 23,818.5715 |
| Neyman Square Root Allocation | 13,354.2962 | 7,482.3705 | 20,836.6667 |

The variance estimator from the stratified random sampling design is:

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

where $h = 1{,}2$ and $\rho_{xy} = 0.7670$ and $S_h^2$ is the stratum variance of the residuals $e_{dk}$ where $e_{dk} = y_{dk} - x_k^T \hat{\beta}_d$.

The optimum value of $n_h$ for the Neyman allocation gave the minimum variance sought. The results of the design variance estimation are presented in table 3.

**Table 3**: Variance Estimation

| Stratum | $N_h$ | $n_h$ | $N_h - n_h$ | $(N_h - 1)\rho$ | $S_h^2$ | $(N_h - 1)\rho(N_h - n_h)\frac{S_h^2}{n_h}$ |
|---|---|---|---|---|---|---|
| 1. | 7,396 | 805 | 6,591 | 5,671.9650 | 0.3262 | 15,148.6151 |
| 2. | 1,553 | 195 | 1,358 | 1,190.3840 | 0.4326 | 3,586.2351 |
| Total | 8,949 | | | | | 18,734.8502 |

### 3.3 Comparison with Global 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; $m = 1,2,..,M (= 2{,}500)$ we define

**i.** $B(\hat{Y}_{d,w}) = \hat{Y}_{d,w} - \hat{Y}_{d,w}^{(m)}$ where $\hat{Y}_{d,w}^{(m)} = \frac{1}{M_d}\sum_{m=1}^{M_d} \hat{Y}_{d,w}^{(m)}$

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

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

**iv.** $AL(\hat{Y}_{d,w}) = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{Y}_{U,d,w}^{(m)} - \hat{Y}_{d,w}\right)^2$

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

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

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 $(\hat{Y}_{U,d,w}, \hat{Y}_{U,d,w})$ is constructed, where

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

where $\hat{Y}_{L,d,w}$ is the lower confidence limit , $\hat{Y}_{U,d,w}$ is the upper confidence limit and $V\left(\hat{Y}^{(m)}_{d,w}\right) = \frac{1}{M_d - 1}\sum_{m=1}^{M_d}\left(\hat{Y}^{(m)}_{d,w} - \overline{\tilde{\bar{Y}}}_{d,w}\right)^2$ .

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,…, M*), a Bernoulli sample is drawn where each unit is selected into the sample independently, with inclusion probability $\pi_k = \frac{N_h}{n_h}$ 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 $\left(\hat{Y}_{L,d,w}, \hat{Y}_{U,d,w}\right)$ 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 |
| $\hat{Y}_{d,w}$ | 0.0074 | 0.0132 | 2587 | 823.23 | 0.768 |

## 4. DISCUSSION

An assisting model of the form $y_{hi} = \beta_0 + \beta_1 x_h + e_h$ where $h$ is the number of strata (domains) and $e_h \sim \mathrm{N}(0, \sigma^2_{\mathrm{e_h}})$. 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 15,148.6151 and 3,586.2351 for stratum 1 and stratum 2 respectively. Similarly, the variance estimate is 18,734.8502. Following from the above estimates, we deduced that the design strata estimates are minimized when the elements of the inclusion probability are not equal but are in the same stratum under calibration approach to domain estimation. We also deduced that design strata estimates sum up to the finite population estimates.

Analysis for the comparison of performance of estimators showed that the biases of 0.74 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 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 $x$.

## 5. CONCLUDING REMARKS

In calibration estimation the common practice is to generate artificial population(s) for simulation study and assign samples to the said population(s) by proxy. We have demonstrated the use of analytical approaches to allocate optimal samples to subpopulations by conducting real data analysis. We recommend analytical approaches for allocation of optimal samples to population(s) or subpopulation(s) through real data analysis as this guarantee the applicability of the proposed estimator(s) to real life situation(s). That is, focus should be on assessing the applicability of the proposed estimator(s) to real life situation(s) through real data analysis rather than on assessing the performance of the proposed estimator(s) against a given estimator(s) through simulation study. Though both cases, could be investigated as it is demonstrated in this article.

## 6. CONCLUSIONS

Calibration estimation for finite population by Deville and Sarndal (1992) is extended to domain estimation in the context of stratified random sampling design. We proposed calibration estimator based on the stratified random sampling design in the presence of non-response. The calibration assumption of reliant on implicit linear relationship between the study variable,$y$ and the auxiliary variable $x$ is retained for the domain estimation.

The problem of optimal allocation of sample sizes for domain estimation has received less attention than merited in the statistical sample survey theory literature. This article equally addressed this problem especially when it is feasible to select sample in every domain and we used the stratified random sampling design (STRS) where domains constitute strata in the sampling design to obtain optimal stratum sample sizes. Six optimal allocation criteria were considered, namely; optimum allocation, Neyman allocation, optimal power allocation, Neyman power allocation, square root allocation and Neyman square root allocation. Analysis showed that among this class of optimal allocation criteria, the Neyman allocation provided the optimal stratum sample sizes that minimized the variance of our proposed calibration estimator.

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 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

Akwa Ibom State Government (2005). Report of the socio-economic study of Akwa Ibom State. Ministry of economic development, Uyo, Akwa Ibom State - Nigeria.

Bankier, M.D. (1988). Power allocation: determining sample sizes for subnational areas. The American Statistician, 12 (3), 174-177.

Cochran, W.G. (1977). Sampling techniques. New York: Wiley and Sons.

Deville, J.C. & Sarndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.

Gamrot, W. (2006). Estimation of a domain total under nonresponse using double sampling. Statistics in Transition, 7(4), 831-840.

Hartley,H.O.(1959). Analytical studies of survey data. Rome: *Instituto di Statistica.*

Hidiroglou, M.A. & Patak, Z. (2006). Domain estimation using linear regression. Survey Methodology, 30(1), 67-78.

Horvitz,D.G. & Thompson, D.J.(1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663-687.

Torabi,M., Datta,G. & Rao ,J.N.K. (2009). Empirical Bayes estimation of small area means under nested error linear regression model with measurement errors in the covariates. Scandinavian Journal of Statistics,36,355-368.

Udofia, G.A. (2002).Estimation for domains in double sampling for probabilities proportional to size. *Sankhya*, B64,82-89.

Yates, F. (1953). Sampling methods for censuses and surveys. London: Charles W. Griffin.

**[1]E. P. Clement, [2]G. A. Udofia and [2]Ekaette I. Enang**
[1]Department of Mathematics and Statistics University of Uyo, P.M.B.1017
Uyo, Akwa Ibom State – Nigeria.
[2]Department of Mathematics, Statistics and Computer Science
University of Calabar, Calabar, Nigeria
E-mail: *epclement@yahoo.com*

**Received**: 12.12.2013

**Revised**: 20.02.2015