ESTIMATION FOR DOMAINS PARAMETERS IN DOUBLE SAMPLING FOR STRATIFICATION WITH NON-LINEAR COST FUNCTION IN THE PRESENCE OF NON RESPONSE

DAVID ANEKEYA ALILAH

A proposal submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics of Masinde Muliro University of Science and Technology

March, 2016

DECLARATION

This research thesis is my original work prepared with no other than the indicated sources and support and has not been presented elsewhere for a degree or any other award.

Signature..... Date .....

David Anekeya Alilah

SES/H/01/15

APPROVAL

The undersigned certify that they have read and hereby recommend for acceptance of Masinde Muliro University of Science and Technology a research proposal entitled "Estimation for domains in double sampling for stratification with non-linear cost function in the presence of non response".

Signature..... Date.....

Dr. Christopher Ouma Onyango

Department of Statistics and Actuarial Science

Kenyatta University.

Signature..... Date.....

Prof. Kennedy Nyongesa

Department of Mathematics

Masinde Muliro University of Science and Technology.

ABSTRACT

In sample surveys separate estimates of a parameter maybe required for sub-populations into which a population is subdivided without separately sampling from these sub-populations. Such sub-populations are called domains of study. Most studies have been carried out on domain estimation using linear cost function. This study therefore estimates domains parameters in double sampling for stratification in the presence of non-response with non-linear cost function which will minimize cost for a fixed value of variance or minimize variance for fixed cost in the presence of non-response. The optimum stratum sizes of a given set of non-linear unit costs are derived using double sampling for stratification. The relative precision of the estimators are empirically compared with corresponding existing estimators. Data simulation will be done and then analysis and evaluation done using R-package.

TABLE OF CONTENTS

DECLARATION ii

ABSTRACT iii

TABLE OF CONTENTS iv

CHAPTER 1:INTRODUCTION 1

1.1 Introduction 1

1.2 Background of the study 1

1.2.1 Domains 1

1.2.2 Double sampling for stratification 3

1.2.3 Optimal allocation in double sampling for stratification-Linear cost function 3

1.2.4 Optimal allocation with non-linear cost function 4

1.2.5 Estimation of domain characteristic/parameters in stratified sampling design 6

1.2.6 Sample design for calibration estimator 6

1.2.7 Optimal allocation in estimation of domain in stratified sampling 7

1.2.8 Non linear cost function 9

1.3 Statement of the problem 10

1.4 Objectives of the Study 11

1.4.1 Specific Objectives 11

1.5 Significance of the Study 11

CHAPTER 2:LITERATURE REVIEW 12

2.1 Sample allocation 12

2.2 Domain estimation 13
2.3 Summary 16

CHAPTER 3:METHODOLOGY **17**

3.1 Introduction 17
3.2 Methods 17

REFERENCES **18**

WORK PLAN 22
BUDGET 23

CHAPTER 1

INTRODUCTION

1.1 Introduction

In most sampling situations information on the variables that defines the domains of study is not either readily available before the start of survey or units in the domain may be identified before sampling is done. However, the estimates of parameters of interest are usually required for subclasses of the population under study which are domains of study. The exact total domain units in each constituent stratum are unknown whenever the domains of study are unknown or they are treated as separate stratum from which a specific sample maybe taken if the domains are known before sampling. In chapter one of this proposal sections 1.2.1 gives a brief description of domains, 1.2.2 double sampling for stratification, 1.2.3 optimal allocation in double sampling for stratification with linear cost function, 1.2.4 optimal allocation with a non- linear cost function 1.2.5 domain estimation in stratified sampling, 1.2.6 sample design calibration estimator, 1.2.7 optimal allocation in estimation of domains in stratified sampling 1.3 statement of the problem, 1.4 objectives of the study and 1.5 significance of the study. Chapter two enumerates some of the related literature and the available gaps. Finally chapter three gives the methods to be employed in order to achieve the objectives.

1.2 Background of the study

1.2.1 Domains

Domain is a subgroup of the whole target population of the study for which specific estimates are needed. In sampling, estimates are made for each of as number of classes into which the population are subdivided; for instance, the focus may not only be the unemployment rate of the entire population but also the breakdown

by age, gender, and education level. Units of domains may sometimes be identified priori to sampling. In such cases the domains can be treated as separate stratum from a specific sample maybe taken. Stratification ensures a satisfactory level of representativeness of the domains in the final sample. These domains are called planned domains.

The precision threshold and or minimum effective sample sizes are set up for effective planned domains. The minimum sample sizes required to achieve a relative margin error of 100.k% for the total Y_d (Domain total) of a study variable y over domain U_d of size N_d given by

$$n_d(min) = \frac{Z^2_{1-\alpha/2} \cdot N_d^2 S^2_{y_d}}{K^2 Y_d^2 + Z^2_{1-\alpha/2} N_d S^2_{y_d}} \tag{1.1}$$

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

For unplanned domain for which units cannot be identified prior to sampling, the need for estimates of certain domains is often evident only after the sampling design has been decided or after the sampling and field work have been completed. The sample sizes for sub-populations are random variables since formation of these sub-populations is unrelated to sampling design. Hence, the size of unplanned domain cannot be controlled. The random size of the sample builds an additional component of variability into the domain estimates.

Consider a random sample S of size n selected without replacement from target population U, of size N. Let S_d be the part of size n_d of the whole sample S which

falls into a domain $U_d(U_d \subseteq U)$, n_d is a random sample which satisfies the following properties.

$$\begin{aligned} E(n_d) &= nP_d \\ V(n_d) &= nP_d(1-p_d) \end{aligned} \tag{1.2}$$

where $P_d = \frac{N_d}{N}$ is the relative size of the domain U_d in the population U.

1.2.2 Double sampling for stratification

In double sampling population is first stratified into H strata (classes). The first sample is a simple random sample of size n' selected from the whole population. Units belonging to a particular stratum and categorized into strata, say $n'_1, n'_2, \ldots\ldots, n'_h, \ldots, n'_H$ such that

$$n' = \sum_{h=1}^{H} n'_h$$

In the second phase n_h units are selected from n'_h such that

$$n = \sum_{h=1}^{H} n_h$$

. In selecting n_h units which are random subsamples of n'_h we select n_h such that $n_h = V_h n'_h$ where $0 < V_h < 1$.

The objective of the first sample is to estimate the strata weights and birth of the second sample is to estimate the strata means $\hat{\bar{y}}_h$ The estimate of the population mean is given by

$$\begin{aligned} \hat{\bar{Y}} &= \sum_{h=1}^{H} w_h \bar{y}_h, \quad w_h = \frac{n'_h}{n'}, \quad \bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{h_i} \\ E(\hat{\bar{Y}}) &= \bar{Y} \end{aligned} \tag{1.3}$$

The estimate of the population variance is given by;

$$\begin{aligned} Var(\hat{\bar{Y}}) &= V_1 E_2(\hat{\bar{Y}}) + E_1 V_2(\hat{\bar{Y}}) \ldots \\ &= \frac{N-n'}{N}\frac{S^2}{n'} + \sum_{h=1}^{H} W_h \frac{S_h^2}{n'}\left(\frac{1}{V_h} - 1\right) \end{aligned} \tag{1.4}$$

The standard way of obtaining the estimate of variance is given by;

$$Var(\hat{\bar{Y}}) = \frac{N-1}{N}\sum_{h=1}^{H}\left(\frac{n_h'-1}{n'-1} - \frac{n_h-1}{N-1}\right)\frac{w_h S_h^2}{n_h} + \frac{N-n'}{N(n'-1)}\sum_{h}^{H} w_h(\hat{\bar{y}}_h - \hat{\bar{Y}})^2 \quad (1.5)$$

1.2.3 Optimal allocation in double sampling for stratification-Linear cost function

The objective is to choose n' and v_h so as minimize $V(\hat{\bar{y}})$ for a specified cost. Let c' be the cost of classification per unit and c_h the cost of measuring a unit stratum h. The cost function is given by

$$C = c'n' + \sum_{h_1}^{H} c_h n_h \quad (1.6)$$

since n_h are random variables we minimize the expected cost for chosen n' and v_h

$$E(C) = C^* \quad = \quad c'n' + n'\sum_{h_1}^{H} c_h v_h W_h \quad (1.7)$$

and variance is given by

$$= \quad n'(V + \frac{S^2}{N}) = (S^2 - \sum_{h=1}^{H} W_h S_h^2) + \sum_{h=1}^{H}\frac{W_h S_h^2}{v_h} \quad (1.8)$$

By use of Cauchy Schwartz inequality,

$$v_h = S_h\left[\frac{c'}{c_h(S^2 - \sum_{h=1}^{H} W_h S_h^2)}\right]^{\frac{1}{2}} \quad (1.9)$$

By substitution of the optimum v_h the minimum variance is found to be

$$V_{min}(\hat{\bar{Y}}) = \frac{1}{C^*}\left[\sum_{h=1}^{H} W_h S_h\sqrt{C_h} + (S^2 - \sum_{h=1}^{H} W_h S_h^2)^{\frac{1}{2}}\sqrt{c'}\right]^2 - \frac{S^2}{N} \quad (1.10)$$

1.2.4 Optimal allocation with non-linear cost function

Optimal sample allocation involves determining the sample size $n_1, n_2, \ldots, n_H$ that minimizes the various characters under the given sampling budget C(where C is the upper limit on the total cost of the survey). The linear cost function

$$C = c'n' + \sum_{h_1}^{H} c_h n_h$$

is appropriate when the major cost item is that of taking the measurement on each unit. If travel costs between units are substantial, empirical and mathematical studies suggest that the travel costs are better represented by the expression

$$\sum_{h=1}^{H} t_h \sqrt{n_h} \tag{1.11}$$

where t_h is the travel cost incurred in enumerating a sample unit in the h^{th} stratum. It's observed that the shortest distance among k randomly scattered points/sampling units is asymptotically proportional to$\sqrt{k}$ for large k

The linear cost function used in stratified sampling in the case of non-response is given by

$$c_h = c_{h_0} n_h + c_{h_1} n_{h_1} + c_{h_2} n_{h_2} \tag{1.12}$$

For large sample sizes the cost function is given by

$$C = c_0 + \sum_{h=1}^{H} c_h + \sum_{h=1}^{H} t_h \sqrt{n_h} \tag{1.13}$$

where t_h is travel cost for a unit within the h^{th} stratum.

Assuming the above non-linear cost function

$$\sum_{h=1}^{H} t_h \sqrt{n_h} + t_0 \leq C$$

Where t_0 is the overhead cost. The restrictions on the sample sizes from various strata are;

$$2 \leq n_h \leq N_h$$

The allocation problem with no non-linear cost function can now be written as

$$V_{min} = \sum_{h=1}^{H} \frac{W_h^2 S_{h_j}^2}{n_h} \tag{1.14}$$

subject to

$$\sum_{h=1}^{H} t_h \sqrt{n_h} + t_0 \leq C$$

Where

$$S_{h_j}^2 = \frac{1}{N_h - 1} \sum_{h=1}^{N_h} (y_{h_{ij}} - \bar{Y}_{hi})^2$$

is the variance of the j^{th} character in the h^{th} stratum

1.2.5 Estimation of domain characteristic/parameters in stratified sampling design

Consider the finite population under study U of size N divided into D domains; $U_1, U_2, \cdots, U_D$ respectively. Domain membership of any population unit is unknown before sampling. Its assumed the domains are quite large, for a typical d^{th} domain, u_d, several characteristics maybe defined including the domain total.

Domain total

$$Y_{u_d} = \sum_{u_d} y_{d_k}$$

Domain mean

$$\bar{Y}_{u_d} = \frac{1}{N_d} \sum_{u_d} y_{d_k}$$

Domain variance

$$S_{u_d}^2(Y) = \frac{1}{N_d - 1} \sum_{k \in u_d} (Y_d - \bar{Y}_{u_d})^2$$

Domain covariance between two characteristics X and Y

$$Cov_{u_d}(X, Y) = \frac{1}{N_d - 1} \sum_{k \in u_d} (x_{d_k} - \bar{X}_{u_d})(y_{d_k} - \bar{Y}_{u_d})$$

1.2.6 Sample design for calibration estimator

For a stratified random sampling design with H strata and n_h elements from N_h in stratum h; $h = 1, 2, \ldots, H$, the design weights needed for the point estimation are

$$d_k = \frac{N_h}{n_h}$$

for all k in stratum h, $k = 1, 2, \ldots, N_h$.

Thes 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 the equation

$$\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(\frac{d_k d_l}{d_{kl}} - 1\right)E_k E_l$$

$$\begin{aligned}
\hat{V}_p(\hat{Y}_d, w) &= \sum_{h=1}^{H}\sum_{k=1}^{N_h}\left\{\frac{\left(\frac{N_h}{n_h}\right)^2\left(\frac{N_h-1}{n_h-1}\right) - \frac{N_h}{n_h}\left(\frac{N_h-1}{n_h-1}\right)}{\frac{N_h}{n_h}\left(\frac{N_h-1}{n_h-1}\right)}\right\}E_k E_l \qquad (1.15)\\
&= \sum_{h=1}^{H}\sum_{k=1}^{N_h}\frac{N_h}{n_h}\left[\frac{N_h - n_h}{N_h}\right]E_k E_l\\
&= \sum_{h=1}^{H}\frac{N_h^2}{n_h}\left[\frac{N_h - n_h}{N_h}\right]E_k E_l\\
&= \sum_{h=1}^{H}N_h^2\left(\frac{1-f}{n_h}\right)E_k E_l \qquad (1.16)
\end{aligned}$$

The variance of the estimator is

$$V_p(\hat{Y}_d, w) = \quad = \sum_{h=1}^{H}N_h^2\left(\frac{1-f}{n_h}\right)Cov(e_k e_l) \qquad (1.17)$$

where $Cov(e_k e_l) = \sigma_h^2\rho$ and from the principle of stratified random sampling

$$\sigma^2 = \left(\frac{N-1}{N}\right)S^2$$

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

$$Cov(e_k e_l) == \left(\frac{N_h - 1}{N_h}\right)S_h^2\rho$$

1.2.7 Optimal allocation in estimation of domain in stratified sampling

The optimum $n(n_h$ optimum) that minimizes the variances of proposed calibration estimators for a specified costs or that minimizes the cost for a specified variance can be considered using simple linear sampling cost function of the form

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

where c_0 is the overhead costs, and c_h is the cost per unit of obtaining the necessary information in the h^{th} stratum. Further, we consider the following allocation forms.

(i) Optimum allocation

Using the cost function $C = c_0 + \sum_{h=1}^{H} c_h n_h$ and the corresponding Lagrangian as

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

By taking the derivatives with respect n_h and λ and equating to zero will finally obtain a solution of n_h as

$$n_{h,opt} = \frac{(C - c_0) S_h \frac{\sqrt{N_h(N_h-1)}}{\sqrt{c_h}}}{\sum_{h=1}^{H} c_h S_h \frac{\sqrt{N_h(N_h-1)}}{\sqrt{c_h}}} \tag{1.18}$$

(ii) Neyman Allocation

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

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

(iii) Optimum power allocation

Let the loss function be

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

the corresponding Lagrangian is

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

$$+\lambda \left\{ \sum_{h=1}^{H} c_h n_h + c_0 - C \right\}$$

Taking the partial derivatives of G_L with respect to n_h and λ respectively and equating to zero and solving for both λ and n_h respectively, we obtain

$$n_{h,opt} = \frac{(C - c_0) S_h N_h^p \frac{\sqrt{N_h(N_h-1)}}{\sqrt{c_h}}}{\sum_{h=1}^{H} c_h S_h N_h^p \frac{\sqrt{N_h(N_h-1)}}{\sqrt{c_h}}} \tag{1.19}$$

(iii) Neyman power allocation

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

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

(iv) Squareroot allocation

If the value of the power allocation is set to be one half (ie 0.5) then

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

(v) Neyman Squareroot allocation

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

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

1.2.8 Non linear cost function

Optimal sample allocation involves determining the sample size $n_1, n_2, \ldots, n_H$ that minimizes the variance of variables characters under the given sampling budget C

(where C is the upper limit on the total cost of the survey) within any stratum. The linear cost function is appropriate when the major item of cost is that of taking the measurement on each unit.

If travel cost between units in a given stratum are substantial empirical, mathematical studies indicate that the costs are better represented by the expression

$$\sum_{h=1}^{H} t_h \sqrt{n_h}$$

where t_h is the travel cost incurred in enumerating a sample unit in the h^{th} stratum.

Beardwood *et al.*, (1959) observed that the distance between k randomly scattered points is proportional to $\sqrt{k}$. Assuming this is non-linear cost function one should have

$$\sum_{h=1}^{H} t_h \sqrt{n_h} + t_0 \leq C$$

where t_0 is the overhead cost. The restriction on the sample strata is

$$2 \leq n_h \leq N_h$$

The allocation problem with nonlinear cost function can now be written as

$$V_{min} = \sum_{h=1}^{h} \frac{W_H^2 S_{h_j}^2}{n_h}$$

subject to

$$\sum_{h=1}^{H} t_h \sqrt{n_h} + t_0 \leq C$$

$$S_{h_j}^2 = \frac{1}{N_h - 1} \sum_{h=1}^{N_h} (y_{hji} - \bar{Y}_{hj})^2$$

is the variance of the j^{th} character in the h^{th} stratum.

In many practical situations, the travel costs t_h in the various stratum are not fixed and maybe considered as random. Let us assume that t_h where $(h = 1, 2, \ldots, H)$ are independently random variables. Let $t' = (t_1, t_2, \ldots, t_H)$ and $n' = (n_1, n_2, \ldots, n_H)$. Then the function $t'\sqrt{n} + t_0$ will also be normally distributed with

mean

$$E\left(\sum_{h=1}^{H} t_h\sqrt{n_h} + t_0\right) = \sum_{h=1}^{H} \sqrt{n_h}E(t_h) + t_0 = \sum_{h=1}^{H} \mu_h\sqrt{n_h} + t_0$$

and variance

$$V\left(\sum_{h=1}^{H} t_h\sqrt{n_h}\right) = \sum_{h=1}^{H} n_h V(t_h) = \sum_{h=1}^{H} n_h\sigma_h^2$$

1.3 Statement of the problem

A number of studies have been carried out on domain estimation. In all these studies the assumption has been that the cost function is linear.However with distances among a given number of randomly selected study points being asymptotically proportional to the square root of the number of points, the aspect of non-linear cost function arises. Moreover the inclusion of the travel cost for measuring a unit within a stratum rules out the assumption of linearity. This study therefore focuses on the estimation of domain parameters with non-linear cost function(which is the case in practice)in the presence of non-response with the aim of minimizing variance for a specified cost or minimizing cost for specified variance.

1.4 Objectives of the Study

To estimate the domain totals, means and variances with non-linear cost function in the presence of non-response

1.4.1 Specific Objectives

In this research we intend to:

(i) To derive the formula for obtaining optimum stratum sample sizes for a given set of unit cost of sample using double sampling for stratification with non-linear cost function in the presence of non-response

(ii) To estimate the domain parameters using double sampling for stratification with non-linear cost function which minimizes cost for a specified value of

variance or minimized variance for a specified value of cost .

(iii) To compare empirically the relative precision of the derived estimators with existing ones.

1.5 Significance of the Study

In many human surveys domains information is in most cases not obtained from all the units in the survey. Even if obtained, its expensive to carry out surveys for individual domains besides variable of interest. The cost implication may not be linear as the distances traveled in classifying the domains seems non-linear. The question of precision with minimum cost and the cost at minimum variance is inevitable. The study will therefore establish the most precise method of estimating domains and most effective allocation methods when the cost implications are not linear.

CHAPTER 2

LITERATURE REVIEW

2.1 Sample allocation

Tschuprow [20] and Neyman [17] proposed the allocation procedure that minimizes variance of sample mean under a linear cost function of sample size $n = \sum_{h=1}^{H}$ in stratified random sampling. Neyman [17] used Lagranges multiplier optimization technique to get optimum sample sizes for a single variable under study.

Beardwood *et al.* [2] came up with quadratic cost function with measurement unit cost and traveling cost within a strata. He proposed that the shortest route among k randomly allocated sampling units in the region is asymptotically proportional to the $\sqrt{k}$ for a large k.

Cochran [6] noted that it is difficult to work out an allocation of sample size which is optimum for all characteristics unless the characteristics are highly correlated and the variations between the stratum variables is very small. Compromise allocations is based on such criteria.

Bankier [1] proposed a "power allocation" as a compromise between Neyman allocation and equal allocation. According to Bankier, if we let

$$C_h = \frac{s_h}{\bar{Y}_h}$$

be the stratum coefficient of variation the power allocation is given by

$$n_h^B = n\frac{c_h X_h^q}{\sum_h c_h X_h^q}$$

where $h = 1, 2, \ldots, L$, X_h is some measure of size or importance of stratum, h and q is a turning constant. The power allocation is obtained by minimizing

$$\sum_h \{X_h^q CV(\bar{y}_h)\}^2$$

subject to

$$\sum_h n_h = n$$

where $CV(\bar{y}_h)$ is the coefficient of variation of the stratum sample mean $\bar{y}_h$. If $q = 1$ and $X_h = N_h\bar{Y}_h$ in the Bankier allocation equation leads to Neyman allocation. Chernyak [4] developed the method of defining optimum sampling fraction among non respondents with non-linear cost function which minimizes costs for a fixed value of variance.

Holmberg [13] addressed the compromise allocation in multivariate stratified sampling problem by taking into consideration the minimization of some of the variances or coefficients of variations of population parameters and minimization of some of the efficiency loss which may result due to increase in variance due to the use of compromise allocation.

Costa *et al* [7] proposed a compromise allocation based on convex combination of proportional allocation $n_h = nW_h$ and equal allocation $n_h = \frac{n}{L}$, where n_h is the stratum sample size, n sample size, L is the number of strata and $W_h = \frac{N_h}{N}$ is the relative size of the of the stratum.

Longford[16]made a systematic study allocation in stratified simple random sampling by introducing "inferential priorities" P_h for the stratum h and G for the population. In particular he assumed that $P_h = N_h^q$ for specified $q(0 \leq q \leq 2)$. He also considered small strata sample size n_h in which composite estimators of strata means $\bar{Y}_h$ maybe used.

2.2 Domain estimation

Yates [23] 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 number of domain elements falling in a random sample of a fixed size, unknown before the start of surveys.

Hartley [10] gave a derivation of Yates results in multistage sampling. His paper was one of the first attempt to unify the theory of domain estimation. He provided theories for a number of sample designs where domain estimation was of interest. He came up with estimation that did not make use of auxiliary information. He also considered a case of ratio estimation where the population totals were known for the domains.

Durbin [9] supported the use of conditional inference to do comparison or choosing the best estimators. To quote him he stated, "If the sample size is determine by a random mechanism and one happens to get a large sample, then one knows perfectly well that the quantities of interests are measured more accurately than they would have been if the sample size happened to be small. It seems self evident that one should use the information available on sample size in the interpretation of the results. To average over variation in the sample size which seemed to have occurred, but did not occur, when the sample size is exactly known, seems quite wrong from the analysis point of view of data actually observed."Kish [15] considered allocation of resources when domain of study are of primary interest.

Holf and Smith [12] considered conditional inference and applied it to study the properties of post stratified estimators in simple random sampling. Rao [18] introduced the idea of "recognizable subsets" of the population to formalize the conditioning process. Recognizable subsets are defined after the sample has been drawn. In the context of domain estimation, the number of units belonging to a particular domain is a random variable. Recognizable subsets in this context are those units in which the sample sizes are fixed within a given domain.

Bethel *et al.* [3] came up with non-linear cost allocation in estimation of population means of multiple variables under stratified random sampling. Deville and Sarndal [8] introduced calibration as a tool for re-weighting non response for the estimation of finite population characteristics like means, ratios and totals. In their study, they found out that the calibration approach requires the formulation of suit-

able auxiliary variables. The calibration approach provides a unified treatment in the use of auxiliary information in the surveys with non response. They observed that in the presence of a powerful auxiliary information the calibration approach meets the objective of reducing both sampling error and the non-response error.

Hidiroglou and Patak [11] in their paper on the domain estimation using linear regression studied the properties of a number of domain estimators of totals in the presence of auxiliary data. They found out that post stratified ratio (POSTR) and modified alternate ratio (MODR) estimators performance in terms of unconditional relative mean squared error efficiency and coverage rate was superior to other domain estimators like the Horvitz Thompson estimator.

Udofia [21] in his article considered estimation of domains by double sampling for probabilities proportional to size (PPS) selection method in a population with known-constituent strata. He considered estimators for domain totals that cut across all strata with unknown weights and made comparison with corresponding global estimators.

Torabi[19] proposed an empirical Bayes estimation of domain means under nested error regression model with measurement errors in co-variates. Varshney *et al* [22] came up with a quadratic cost function for a large sample size as

$$C = c_0 + \sum_{h=1}^{L} c_h + \sum_{h=1}^{L} \tau_h \sqrt{n_h}$$

where τ_h is the travel cost for enumerating a unit within the h^{th} stratum.For the case of non-response, this equation was further simplified to

$$\sum_{h=1}^{L} (t_{h_0} + t_{h_1} W_{h_1}) \sqrt{n_h} + \sum_{h=1}^{L} t_{h_2} \sqrt{u_n} \leq c_0$$

where$t_{h_1}$is travel cost for the respondents unit within the h^{th} stratum and $u_h$is the sub-sample from non-respondents units

Choudhry[14] considered sample allocation issues in the context of estimating sub-populations(stratum and domain) means as well as the aggregate population

means under stratified simple random sampling. In his method non-linear programming was used to obtain 'optimal' sample allocation to strata that maximizes the total sample sizes subject to a specified tolerances on the coefficient of variation estimators of the strata and population means. The resulting sample size was used to determine the sample allocations for the methods of Costa *et al* [7] based on compromised allocation and of Longford[16] specified based on "inferential properties". They also came up with the idea on sample allocation to strata when the reliability requirements for domains cutting across strata are specified. They concluded that the non-linear programming(NLP) method of sample allocation to strata under stratified random sampling minimizes the total sample size subject to specified tolerances on the coefficient of variation of estimators both the strata and population means.

Clement *et al* [5] developed analytical approach for finding the best sampling design subject to a cost constraint. They considered stratified random sampling design where elements of the inclusion probabilities are not equal but are in the same stratum and proposed estimators of the totals for domains of study under non-response in the context calibration estimators. In their analysis they found that Neyman allocation provided the optimal stratum sample sizes that minimized the variance of the proposed calibration estimators.

2.3 Summary

In summary from all these studies its evident that the researchers have confined themselves to estimation of domains with the assumption that the cost function is linear. Clement *et al* [5] and Udofia [21] who recently have led remarkable steps in the studies on domain estimation, have made assumption that the cost function is linear. More so Clement *et al* [5] in developing analytical approach to best sampling design considered stratified sampling without inclusion of travel cost for measuring a unit in a stratum which is an important attribute of cost function.This study

therefore aims at developing double sampling for stratification with a non-linear cost function as an efficient design in estimating domains in comparison with other recently developed design methods putting in mind the non-linear and travel costs components.

CHAPTER 3

METHODOLOGY

3.1 Introduction

To achieve our objectives the following methods will be employed

3.2 Methods

(1) A model for estimation of domain parameters using double sampling for stratification with non-linear cost function will be derived. In so doing n_h is chosen to minimize variance for a specified cost or minimize cost for specified variance. To achieve this a non-linear polynomial cost function of the form

$$C = c'(n')^{\theta} + \sum_{h=1}^{H} c_h n_h + \sum_{h=1}^{H} t_h \sqrt{n_h}$$

where C is the total cost, c' is the cost of classification per unit, c_h is the cost of measuring a unit per stratum, t_h is the travel cost for measuring a unit within the h^{th} stratum, and n_h is the number of sample units selected from h^{th} stratum.

Using Lagrange's multiplier optimal allocation equations will be derived with non-linear cost function with a positive and negative θ and logarithmic non-linear cost function.

(2) Data collection For the purpose of empirical illustration sample data of 2009 Census from the twelve sub-counties of Kakamega County will be used. Using the data a population with N=30 First Stage Units (FSU) and H=20 second stage units will be generated by combining the adjacent 10 units and allocating them to the respective first stage units. Data will be simulated and used to obtain numerical estimates using each of the above optimal allocation options.

(3) The percentage reduction in the expected cost will be computed as well as the optimum values of sample sizes of different estimators in respect of the controlled variables.

(4) The results will be empirically compared to obtain the relative performance of the proposed estimators with corresponding global estimators using bias, Relative bias, mean square error, the average length of confidence interval, and the coverage probability of the estimates. The analytical studies will be carried out using R statistical package.

REFERENCES

[1] Bankier, M., (1988). Power allocation Determining sample sizes for subnational areas, *The American Statistician*, **42** 174-177.

[2] Beardwood, J., Halton, J.H., and Hammersley, J.M., (1959).The shortest part through many points. *Mathematical proceedings of the Cambridge philosophical society.*, **55** pp. 299-327.

[3] Bethel J., (1989). Sample allocation in Multi-variate surveys, Survey methodology **15** pp. 47-57.

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

[5] Clement E.P, (2014). Udofia G.A.,and Enang E.I, Estimation for domains in Stratified Sampling Design in the presence of non-response, *American Journal of Mathematics and Statistics* **4(2)** pp. 65-71.

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

[7] Costa, A., Satorra, A., and Ventura E., (2004). Using Composite estimators to improve both domains and total area estimation. *SORT*. **28** pp. 69-86.

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

[9] Durbin J., (1958). Sammpling theory for estimates based on fewer individuals than the number selected. *Bulletin of international statistical institute* **36**, pp. 113-119.

[10] Hartley, H.O., (1959) *Analytical studies of survey data*. Rome: Instituto Di Statistica.

[11] Hidiroglou, M., and Patak, Z., (2001). "Domain estimation using linear regression", proceedings of the annual meeting of the American statistical association, August 5-9.

[12] Holf D., and Smith T.M.F., (1979). Post stratification, *Journal of the Royal Statistics Society, sec, A-142*, pp. 33-46.

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

[14] Hussain Choudhry G., Rao, J.N.K, and Michael A., Hidiroglou, (2012). On sample allocation for efficient domain estimation,*survey methodology, June 2012, Vol 38, No. 1, pp. 23-29, Statistics Canada*, Catalogue No. 12-001.

[15] Kish, L., (1961). Efficient allocation of a multipurpose sample.*Econometrica* **29** pp. 363-389.

[16] Longford, N. T., (2006). Calculation of small area estimation survey methodology. **32** pp. 87-96.

[17] Neyman C. and Jerzy D. (1934).; On the two diherent aspects of the representative methods of stratified sampling and the method of purposive selection,*Journal ofg royal statistical society.* **97(4)** pp. 558-625.

[18] Rao, J.N., (1985). Conditional inferences in survey sampling, *Survey Methodology*.pp.15-32.

[19] Torabi, M., Datta, G., and Rao J.N.K., (2009). Emperical Bayes estimation of small area means under nested error linear regression model with measurement errors in the covariates. *Scandivanian Journal of Statistics* **36** pp. 355-368.

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

[21] Udofia G.A., (2002). Estimation for domains in double sampling for probabilities proportional to size. *The Indian Journal of Statistics* **63** pp. 82-89.

[22] Varshney, R., Najmussehr and Ahsan, M.J. (2012). Estimation of more than one parameters in stratified sampling with fixed budget*Mathematical methods of operation research*. **72(2)**, pp. 185-197.

[23] Yates, F., (1953). *Sampling Methods for censuses and surveys*. London: Charles W. Griffins.

WORK PLAN

ACTIVITY	PERIOD
1.PREPARATORY STAGE:	
(i) Acquisition of research materials	September (2014) – January (2015)
(ii) Literature review	February (2015)– July (2015)
(iii)Definition of problem	August (2015)– September (2015)
(iv) Proposal writing	September (2015)– October (2015)
(v) Proposal defense	October (2015)
2.OPERATIONAL STAGE:	
(i) Research(problem–solving)	November (2015)–February (2016)
(ii) Drafting research report	March (2016)– September (2016)
(iii) Revising the draft report	October (2016)– December (2016)
3.EVALUATION STAGE:	
(a) Submission and evaluation of thesis	January (2017)– April (2017)
(b) Thesis defense	May (2017)

BUDGET

Year	Items/Descriptions	Unit cost(Ksh.)	Total(Ksh.)
Year 1	1. Stationary:10 reams of Foolscaps	500	5000
	2. Statistical software; SPSS,Genstat, SAS		250,000
	3. Acquisition of 5 Journals per year	20,000	100,000
	4. Attending 3 seminars and Conferences	50,000	150,000
	SUBTOTAL 1	-	505,000
Year 2	1. Data processing and analysis: Stationary:10 reams of Foolscaps	500	5,000
	2. Attending seminars/conferences (3)	50,000	150,000
	SUBTOTAL 2	-	155, 000
Year 3	1. Thesis preparation and submission: Stationary:10 reams of Foolscaps	500	5,000
	2. Attending seminars/conferences (3)	50,000	150,000
	SUBTOTAL 3	-	155,000
	GRAND TOTAL	-	815,000