

**SAMPLING ERRORS IN SOCIOECONOMIC
SURVEYS FROM RURAL SETTINGS
IN ETHIOPIA**

A Thesis
Presented to the
School of Graduate Studies
Addis Ababa University

In Partial Fulfilment of the
Requirements for the Degree of Master of Science
in Statistics



By

ZELEALEM TAREKEGN

June, 1994

Zel
Sta
1994

To the memory of my father,

TAREKEGN EMIRU

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES



SAMPLING ERRORS IN SOCIOECONOMIC SURVEYS
FROM THE RURAL SETTING IN ETHIOPIA

By

Zeleealem Tarekegn
Department of Statistics
Faculty of Science

Approved by the Examining Board:

Prof. Asmerom Kidane
Chairman, Dept.
Graduate Committee

A handwritten signature in black ink, appearing to read "Asmerom Kidane", written over a horizontal line.

Prof. Asmerom Kidane
Advisor

A handwritten signature in black ink, appearing to read "Asmerom Kidane", written over a horizontal line.

Dr. Igor Litvin
Ext. Examiner

A large, stylized handwritten signature in black ink, written over a horizontal line.

Dr. Eshetu Woncheko
Examiner

A handwritten signature in black ink, appearing to read "Eshetu Woncheko", written over a horizontal line.

ACKNOWLEDGEMENTS

I wish to express my sincere and heartfelt gratitude to my advisor, Professor Asmerom Kidane, for suggesting the topic of this thesis, providing important materials and for his unlimited assistance and encouragement throughout my study.

Many thanks are also due to Ato Yohannes Hilawi who contributed in providing and organizing the data the way that I wanted.

My heartfelt appreciation and thanks are also due to Ato Temesgen Zewotir for his unlimited assistance and encouragement.

I am particularly indebted for invaluable help and encouragement to my wife, Atiteged Alemu.

I would also like to extend my thanks to the Ministry of Education, for giving a full sponsorship and opportunity to join the Graduate School of Addis Ababa university.

Finally, I would like to thank the staff members of the Department of Statistics, my family, and all my friends for their special assistance and encouragement.

ABSTRACT

In this paper an attempt was made to study sampling errors from socioeconomic variables of three different areas of Gondar region under three different sampling techniques. On the basis of the variances (standard errors) of each of the variables, design effects and rate of homogeneity values were computed. Depending on these values of sampling errors the accuracy and efficiency achieved in different sample designs on qualitative and quantitative socioeconomic variables were compared. Looking at the pattern of the sampling errors, out of the three techniques Stratified random sampling was found to be highly superior for both discrete and continuous variables, in all of the three areas. On the other hand these computed values of sampling errors showed that for continuous variables, the performance of Cluster sampling was much better than that of Simple random sampling.

C O N T E N T S

	Page
1. INTRODUCTION	2
1.1 Statement of the problem	2
1.2 Use of sampling errors.....	4
1.3 The data and the survey area.....	6
2. METHODS OF ANALYSIS.....	7
2.1 Various measures of sampling errors.....	7
2.1.1 The design effect.....	7
2.1.2 The Intraclass Correlation (ROH).....	9
2.1.3 Different Sample Designs and Statistics used	12
2.1.3.a Cluster Sampling.....	12
2.1.3.b Stratified Random Sampling.....	19
2.2 Portability of standard errors.....	25
3. ANALYSIS AND RESULTS.....	28
4. CONCLUSION.....	46
REFERENCES.....	49

1. INTRODUCTION

1-1. Statement of the problem

It is known that survey research is aimed at estimating specified population parameters. A population parameter is a numerical expression that summarizes the values of some characteristic(s) for all N elements of an entire population.

Due to different reasons researches on an entire population may be difficult, i.e, due to cost, time and some other related problems. For this reason part of the entire population (sample) is considered. This sample may be selected using different methods of sample designs; to estimate the population parameter by a sample value (statistic) which is computed from the n elements of the sample.

Sample values (statistics) are subject to errors of different types and magnitudes. No matter how scientifically and how carefully a survey may be designed and executed, errors are bound to occur, since only part of the total population is being studied.

We group these errors into two, namely sampling and non-sampling errors. The former resulted due to the sampling process and the latter due to imperfect procedures of observations i.e, inaccuracy of data and incompleteness of cover. Ordinarily sampling errors account for most of the variable errors of a survey. We strive to design a survey so that the sampling error is as small as possible.

This paper deals only with the sampling errors resulting from different types of sample designs. The aim of the project is to have some insights into the degree of reliability and the homogeneity of responses as well as the efficiency of a particular sampling method compared to simple random sampling in the socioeconomic surveys of the rural settings in Ethiopia. Its importance is to compare sampling errors based on various methods of sampling. The two sampling designs that we will be considering are Cluster sampling and Stratified random sampling; Simple random sampling will be a point of reference. In general, the objective of this research is to assess the accuracy and efficiency achieved in Cluster and Stratified sampling in quantitative and qualitative socioeconomic variables compared to Simple random sampling; and to study the variability of standard errors and related statistics with a view of evaluating the degree of precision of the information collected from the surveys.

This study is based on three socioeconomic surveys conducted in the rural settings of Ethiopia, specifically in Gondar region the areas Gondar town, Dabat town and Rural part of Wogera District in the years 1992 and 1993.

The sampling errors for several different variables are computed. For each variable sampling errors for means (proportions) and for differences of means also for diverse subclasses based on the entire sample were computed. Each of the calculations of sampling errors included variances, standard errors (se's), "design effects" and intraclass correlations.

From the computed variances (se's) measures that greatly facilitate the process of ^{imputation} are constructed, namely the design effects and the intraclass correlations. The design effect measures the relative efficiency of the design under use compared with simple random sampling with same sample size considered. For subclasses of diverse sizes values of intraclass correlations were computed; these values of ρ_{oh} (rate of homogeneity) measure the degree to which values for variables are homogeneous within sample clusters. These measures will be discussed in detail in Chapter 2.

1-2. Use of sampling errors

Sampling errors have several uses, from which we will consider the main ones in this paper.

Most socioeconomic surveys are multipurpose and may consist of many questions, some running in hundreds, and one may have to estimate sampling errors for each of the variables under consideration.

If a survey is conducted in different regions of a country or conducted in different countries, one may need to compare the variability in response between any two regions or countries. A socioeconomic survey conducted within a given area may require comparing response variability between subgroups of a given region. All these estimations and comparisons will result in a very large body of sampling errors. This large body of sampling errors needs to be presented in such a way that it is scientific, easy to use and understand. If this is done, the utility and reliability of a survey can be studied by looking

at the pattern and magnitude of the sampling errors. More specifically one is able to identify which questions give more reliable results and which do not.

A second major advantage of sampling errors in socioeconomic and similar surveys is the provision of basic information for the planning and designing of similar surveys for the future. More specifically, if one is interested in acquiring a necessary sample size for a level of precision determined in advance, a carefully prepared and presented set of sampling errors will be essential. This is what Kish, Groves and Krotki, (1975) call "portability" of sampling error, namely the use of sampling error far from its source.

A third and widely known use of sampling errors will be to see how good a particular sampling method fared when compared with simple random sampling. This is estimated by the utilization of the **design effect**, namely the degree of precision of the sampling design under use compared to simple random sampling.

While it is possible to estimate sampling errors and related results for each of the variables as well as the difference between pairs, it is not easy to come out with a single index as an overall measure of sampling error. The second best procedure could be to group the sampling errors by subject matter, by their magnitude or both. In other words, it may be possible to show which topic in the questionnaire (interview) gave good results and which didn't.

If one groups sampling errors by subject or similarity of question, then it may be possible to estimate an "average" standard error. This is better done when sampling errors are

converted into design effect or into an index of homogeneity, which will be discussed in detail in the later sections of Chapter 2.

1-3. The data and the survey area

The data collected in the years 1992 and 1993 by Ato Yohannes Hilawi to study the **Economic Value Of Children** was obtained from Demographic Centre of Addis Ababa University. The survey area covers three parts of Gondar region namely, Gondar town, Dabat town in Wogera District and a Rural area in Wogera District consisting of about ten peasant associations located between Gondar and Dabat towns. The data involves several continuous and discrete socioeconomic variables, out of which some 18 are selected for this study.

The "Kebeles" for Gondar and Dabat towns were used as subdivisions (clusters) while the Rural area was partitioned into "Gotes". As mentioned before the study depends on the three sampling designs; Simple random, Cluster and Stratified sampling.

A simple random sample of households, proportional to the sizes of the areas, were taken from randomly chosen "Kebeles" in the towns; and "Gotes" in the Rural area independently and, similarly in Cluster and Stratified samples simple random sample of "Kebeles" (or "Gotes") were selected independently and the variance (standard error) of the sample mean were computed for each cluster or stratum. For each of the three areas the computed sampling errors under different sampling designs are summarized in the Tables in Chapter 3 of this paper.

2. METHOD OF ANALYSIS

2-1. Various measures of sampling errors

2-1-1. The Design Effect

In almost any survey the sample designer is faced with a number of options for which the evidence does not exist to guide him to an optimal decision. Nearly always he has to take decision on the spot; hardly ever can he conduct supplementary research. Once the survey is completed and the results are in, much more information becomes available but by this time it is too late to do anything about the design. The only helpful strategy is to pass such information onto designers of later surveys. Thus there is a pressing need for a body or knowledge based on survey experiences accumulated in a wide variety of circumstances, upon which the sampling statistician can draw when attempting to optimize his sample design. Hence, the sampling plan for each survey has to be designed, *ad hoc*, in light of various factors, such as what is already known about the population under study i.e, the sampling frame, the availability of maps, the accessibility of certain areas, the members for the field work of the survey, the overall cost and the like. This leads to the concept of the Design Effect or the Design Efficiency Factor (Deff), a measure of the relative efficiency of the design being used as compared with Simple random sampling. Efficiency in this sense is measured by the sampling variance, that is to say, the square of the standard error (se).

This basic statistical concept sets probabilistic limits

to the amount by which the true value under estimate differs from the observed value in the sample. Under certain conditions, the variance can be calculated *a priori* on theoretical grounds; in other cases, it has to be estimated from the data themselves.

A common and convenient criterion asserts that the true value lies within a range of twice the standard error on either side of the sample value. The object of a good design is to reduce this standard error as much as possible so that the population value lies within a short range of the sample value. The Design Efficiency Factor (Deff) is defined as the ratio of the actual sampling variance taking into account the complexity of the sampling design to the variance of the same sample size (n) under assumptions of simple random sample [Kish,1965,8-2].

$$Deff = \frac{V}{V_r} \quad (2.1)$$

where V is the actual sampling variance and V_r is the variance of simple random sampling. If comparisons are made between standard errors (as more often used in this paper) instead of variances the corresponding ratio is denoted by **deft**, where

$$deft = \sqrt{deff} = \sqrt{\frac{V}{V_r}} = \frac{se}{se_r}$$

se = the standard error of the actual sample design,
and se_r = the standard error of the simple random sampling. However, each variable has its own standard error

and consequently the design effect is not an absolute single quantity attached to the design, but is a set of quantities, one for each variable under estimate. One may legitimately speak of the design effect for one variable or the other, but not of the design effect of the whole design unless it is possible to bring together the design effects for all variables of interest into a single index number expressive of the design effect over all the variables.

Using these deff (deft) values we can see whether the sampling method involved is better or worse than simple random sampling.

2-1-2. The Intraclass Correlation [ROH]

In some surveys the sampling units themselves may consist of a group or cluster of smaller units called elements or subunits of the population. Hence the sampling unit in this case is a cluster of elements and not the individual element.

There are two main reasons for the widespread application of cluster sampling. Although the first intention may be to use the elements as sampling units, it is found in many surveys that no reliable list of the elements in the population is available and that it would be very much expensive to construct such a list. So, when individual selection of elements seems too expensive, survey tasks can be facilitated by selecting clusters. When there are no complete and up-to-date lists of people, houses, farms, etc. of any large geographic region, it is possible to divide the region, using its map into areal units such as blocks in cities and segments of land with readily identifiable boundaries in the rural parts.

Even when a list of cluster units is available, due to economic considerations we may go to a choice of a larger cluster unit. For fixed sample size, a small unit usually gives more precise results than a larger unit. But when cost is balanced against precision, the larger unit may prove superior. A rational choice between two types or sizes of units may be made by the familiar principle of selecting the unit that gives the smaller variance for a given cost or the smallest cost for a prescribed variance.

Among different selection procedures, cluster sampling often has by far the greatest effect on both the variance and the cost.

The main factor affecting the design efficiency is the effect of clustering. If, among the group forming the cluster, there are correlations of appreciable size, the amount of information from a set of n individuals is not n times the amount of information derived from one individual.

In practice one may expect some positive correlation between neighbouring members in a cluster and this reduces the efficiency of the sampling design. The central point of interest is how much the efficiency is reduced, considered as a trade-off against the saving in expense achieved by clustering the sample.

A measure of the degree of relationship between the members of a cluster (homogeneity of elements) is the so-called intraclass correlation, which is a kind of average of the correlation among members of a cluster and is usually denoted by the Greek letter rho, ρ .

$\rho = \frac{deff-1}{n-1}$, where deff here denotes the design effect only for

the particular cluster and n the cluster size . When a whole design comprising various clusters is under consideration, an average of the cluster size is taken and we write

$$roh = \frac{deff-1}{\bar{b}-1} \quad (2.3)$$

where, $\bar{b}$ is the average size [Kish, Groves and Krotki, (1976)].

Using roh instead of rho is to remind the reader that it is an average of intraclass correlations and identified in the letters with the initials of Rate Of Homogeneity. If there is no intraclass correlation roh is zero and the deff is unity.

Theoretically roh can attain the value unity, in which case the deff will be high and this means the design is inefficient. In practice values of roh between 0 and 0.2 can be expected. This is to say that, randomly distributed variables have roh values near zero, whereas highly clustered items are found near 0.10 or 0.20, and perfectly segregated variables can theoretically approach 1.0. These were analyzed to search for stable and useful relationships, among them.

2-1-3. Different sample designs and statistics used

2-1-3-a. Cluster sampling

i. The simple mean $\bar{y}$

If a simple random sample of size n elements is selected out of the N elements of a population then, the simple random sample mean is given by:

$$\bar{y} = \frac{Y}{n} = \sum_{j=1}^n \frac{y_j}{n}, \quad \text{with variance}$$

$$\text{var}(\bar{y}) = \frac{1-f}{n} \cdot S^2, \quad (2.4)$$

$$S^2 = \sum_{j=1}^N \frac{(Y_j - \bar{Y})^2}{N-1}$$

where $s^2 = \sum_{j=1}^n \frac{(y_j - \bar{y})^2}{n-1}$, is the sample estimate of S^2 ,

$f = n/N$ and y_j is the j th observation in the sample.

ii. The ratio mean

In most cluster samples, the clusters may have unequal sizes due to different cases. Even a sample designed with equal clusters may most of the time end up with clusters of unequal sizes.

It is trivial that:

- this inequality may exist in the actual sizes of clusters that were designed with equal measure of sizes.
- this inequality may be introduced into the final cluster results due to non responses.
- also if the members of subclasses that cut across the clusters are distributed unequally in the clusters, this

may result in clusters of unequal sizes.

In this paper we will deal with clusters of unequal sizes in the socioeconomic surveys conducted, considering the **ratio mean**.

Let **N** be the total number of individual members grouped into **A** clusters with **N_α** elements in the **α**th cluster.

According to Kish (1965), the sample mean (**ratio mean, r**) and its variance are given as follows:-

$$r = \frac{Y}{X} = \frac{1}{X} \sum_{\alpha=1}^a y_{\alpha} = \frac{\sum_{\alpha=1}^a y_{\alpha}}{\sum_{\alpha=1}^a x_{\alpha}} = \frac{\sum_{\alpha=1}^a \sum_{\beta=1}^{x_{\alpha}} y_{\alpha\beta}}{\sum_{\alpha=1}^a x_{\alpha}} \quad (2.5)$$

where, $y_{\alpha\beta}$ = the β^{th} observation in cluster α

x_{α} = sample size of cluster α

$x = \sum_{\alpha=1}^a x_{\alpha}$ is the overall sample size

$y_{\alpha} = \sum_{\beta=1}^{x_{\alpha}} y_{\alpha\beta}$ is the α^{th} cluster total

$y = \sum_{\alpha=1}^a y_{\alpha}$ is the sample total of the observations

Y is population total of the observations

X is the overall population size.

Note that both x_{α} and y_{α} are random variables because size x_{α} varies from cluster to cluster.

The sample value $r = y/x$ estimates the population value $R = Y/X$. The variance of the ratio mean r , based on a random selection of a clusters is estimated by

$$\text{var}_{cl}(r) = \frac{1-f}{x^2} \cdot \frac{a}{a-1} \cdot [\sum y_{\alpha}^2 + r^2 \sum x_{\alpha}^2 - 2r \sum y_{\alpha} x_{\alpha}] \quad (2.6)$$

An estimate of the variance of the simple mean for a simple random sample of same number of elements, n is as given in Equation (2.4).

Hence,

$$deff = \frac{var_{cl}(r)}{var_{srs}\bar{y}} = \frac{na}{x^2(a-1)s^2} [\sum y_{\alpha}^2 + r^2 \sum x_{\alpha}^2 - 2r \sum y_{\alpha} x_{\alpha}] \quad (2.7)$$

and

$$deft = \frac{se_{cl}(r)}{se_{srs}\bar{y}} .$$

A relatively high variance (se.) of cluster sampling will result in a deff or deft value that is greater than 1. Thus it should be noted that the relative ease and economy of cluster sampling may be compromised with possible loss of precision. This will happen as a result of high degree of homogeneity among the elements within a cluster.

Hence, $roh = \frac{deff-1}{\bar{b}-1}$ and

$$deff = 1 + roh(\bar{b}-1) \quad , \text{ where } \bar{b} \text{ is the average cluster}$$

size. Here we note that the ratio estimates the effect of clustering on sample design in terms of the measure of homogeneity, roh. To be specific this means, if the value of the design effect exceeds 1, cluster sampling has a greater variance which implies that there is a high degree of homogeneity among elements within a cluster. Hence for cluster samples of size B, roh will be indirectly estimated as follows

$$deff = [1 + roh(B-1)] \quad (2.8)$$

or $roh = \frac{deff-1}{B-1} \quad , \quad (2.9)$

and if sub-samples of size b are selected randomly from equal clusters we use

$$deff = \frac{s_a^2/a}{s^2/n} = [1 + roh(b-1)] \quad (2.10)$$

In case of unequal clusters, say α th cluster with size b_α , where $\alpha=1,2,\dots,a$; the average cluster size $\bar{b} = \frac{n}{a}$ is used in

place of b hence,

$$deff = [1 + roh(\bar{b}-1)] \quad (2.11)$$

and

$$roh = \frac{deff-1}{\bar{b}-1}$$

One may regard the formula above (2.10) as measuring the rate of homogeneity, roh in the primary selections of size b as created by the sample design.

Since $b = n/a$, note also that from (2.10)

$$\frac{s_a^2}{s^2} = roh + \frac{1-roh}{b} \quad (2.12)$$

hence we have a measure of precision, relative to the precision of elements, for the created sampling units of size b .

- When $roh > 0$, the $deff$ exceeds 1, because the cluster sample has a greater variance than simple random

sample with the same sample size n .

- _ The highest value of ρ_{oh} is $+1$, this corresponds to, having same value of all the elements comprising any cluster.

In this case $\text{deff} = [1 + \rho_{oh}(b-1)] = b$ and $\frac{s_a^2}{s^2} = 1$,

which implies that variance of cluster units is as large as variance of simple random sampling.

- _ The lowest value of ρ_{oh} is $-\frac{1}{b-1}$ which corresponds to

$[1 + \rho_{oh}(b-1)] = 0$, and to zero variance between cluster means.

- _ Negative values of ρ_{oh} are rare: they occur when cluster means are more uniform than would be produced by random sorting.

- _ If the variable is distributed completely at random among the clusters, then we expect a ρ_{oh} of zero, and the design effect $[1 + \rho_{oh}(b-1)] = 1$. In this case

$s_a^2 = \frac{1}{b} s^2$, the variance of cluster units is $1/b$ as

great as for simple random sampling.

Generally, population variables are not "well-mixed", they are not randomly distributed in groups and clusters. In practical survey clusters, ρ_{oh} tends to be greater than zero, sometimes by large, but quite often by little amounts.

Even a relatively small positive roh value can have a notable effect, $[1 + \text{roh}(b-1)]$, on the variance if the sample cluster size b is large [Kish, 1965].

iii. Difference of two means

Suppose that we wish to estimate the difference of two population means by taking two independent samples from the two different populations. In simple random sampling the estimate of the difference of the sample means $\bar{Y} - \bar{X}$ is denoted by

$\bar{y} - \bar{x}$, and its variance is:

$$\text{Var}(\bar{y} - \bar{x}) = \text{Var}(\bar{y}) + \text{Var}(\bar{x})$$

The sample estimate of this variance is

$$\text{var}(\bar{y} - \bar{x}) = \frac{1-f_y}{n_y} s_y^2 + \frac{1-f_x}{n_x} s_x^2 \quad (2.13)$$

for independent samples, where $f_y = n_y/N$, $f_x = n_x/N$,

$$s_y^2 = \sum_{i=1}^{n_y} \frac{(y_i - \bar{y})^2}{n_y - 1} \quad \text{and} \quad s_x^2 = \sum_{i=1}^{n_x} \frac{(x_i - \bar{x})^2}{n_x - 1} \quad (2.14)$$

Similarly in cluster sampling the difference of the two population ratio means $R = Y/X$ and $R' = Y'/X'$ denoted by $R - R'$ is estimated by $r - r'$, where r and r' are sample cluster means (ratio means) computed from two independent samples of the two populations, and

$$r = \frac{Y}{X} = \frac{1}{X} \sum_{\alpha=1}^a Y_{\alpha} = \frac{\sum_{\alpha=1}^a Y_{\alpha}}{\sum_{\alpha=1}^a X_{\alpha}} = \frac{\sum_{\alpha=1}^a \sum_{\beta=1}^{x_{\alpha}} Y_{\alpha\beta}}{\sum_{\alpha=1}^a X_{\alpha}}$$

$$r' = \frac{Y'}{X'} = \frac{1}{X'} \sum_{i=1}^{a'} Y_{i'} = \frac{\sum_{i=1}^{a'} Y_{i'}}{\sum_{i=1}^{a'} X_{i'}} = \frac{\sum_{i=1}^{a'} \sum_{j=1}^{x_{i'}} Y_{ij'}}{\sum_{i=1}^{a'} X_{i'}}$$

where, the notations for the terms in the first population are as given in section 2-1-3-a and for the second population are as follows

y_{ij}' = the j^{th} observation in cluster i

x_i' = sample size of cluster i

$X' = \sum_{i=1}^{a'} x_i'$ is the overall sample size

$y_i' = \sum_{j=1}^{x_i'} y_{ij}'$ is the i^{th} cluster total

$Y' = \sum_{i=1}^{a'} y_i'$ is the sample total of the observations

Y' is population total of the observations

X' is the overall population size.

The variance of the difference is given by

$\text{Var}(r_{c1} - r'_{c1}) = \text{Var}(r_{c1}) + \text{Var}(r'_{c1})$, since the samples are independent.

The sample estimate of this variance is

$$\text{var}(r_{cl} - r'_{cl}) = \frac{1-f}{x^2} \cdot \frac{a}{a-1} \cdot [\sum y_{\alpha}^2 + r^2 \sum x_{\alpha}^2 - 2r \sum y_{\alpha} x_{\alpha}] +$$

$$\frac{1-f'}{x'^2} \cdot \frac{a'}{a'-1} \cdot [\sum y'_{i'}^2 + r'^2 \sum x'_{i'}^2 - 2r' \sum y'_{i'} x'_{i'}] \quad (2.15)$$

$$\alpha = 1, 2, \dots, a \quad \text{and} \quad i = 1, 2, \dots, a'.$$

Following similar procedures as before, we can solve for the deff (deft) values using

$$\text{deff} = \frac{\text{var}_{cl}(r - r')}{\text{var}_{srs}(\bar{y} - \bar{y}')} \quad (2.16)$$

These values are given in Tables 1a-1, 1b-1 and 1c-1.

2-1-3-b. Stratified random sampling

Stratification can be used to increase the reliability of sample results. Here, the selection of a sample within each stratum is accomplished by simple random sampling of specified sampling units. Ordinarily these sampling units will be the elementary units.

The amount of increase in precision of sample estimates accomplished by stratification will depend on the degree of homogeneity that is achieved within each stratum, or on how much of the variability in the characteristic being estimated is reflected in the difference among the strata. This in turn depends on how effectively the strata have been defined.

The following procedures were followed in the use of stratification and comparisons made with simple random sampling:-

- the entire population of size N sampling units is divided into H distinct sub-populations (strata),
- within each stratum a separate sample is selected from all the sampling units comprising that stratum,
- from the sample obtained from each stratum, a separate stratum mean is computed; the strata means are properly weighted to form a combined estimate for the entire population and finally
- the variances of the means are also computed separately within each stratum and then properly weighted and added into a combined estimate for the population.

Notations:-

N = entire population size

H =total number of strata

N_h = the size of the h^{th} stratum

Y = total sum of the observations in all H strata

$\bar{Y} = \frac{Y}{N}$ =the population mean

Y_h = the h^{th} stratum total

$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi} = \frac{Y_h}{N_h}$ = h^{th} stratum mean of the N_h elements

Y_{hi} = i^{th} observation in stratum h

n_h = number of elements selected into the sample from stratum h

$$n = \sum_{h=1}^H n_h = \text{entire sample size}$$

$W_h = N_h/N = h^{\text{th}}$ stratum weight

$$\bar{Y}_w = \sum_h W_h \bar{Y}_h = \sum_h \frac{N_h}{N} \cdot \bar{Y}_h = \frac{1}{N} \sum_h Y_h = \frac{Y}{N} = \bar{Y} ,$$

$\bar{Y}_w = \bar{Y}_{st}$ is the weighted mean

y_h = the h th stratum total in the sample

$\bar{y}_h$ = the h th stratum sample mean

y_{hi} = the i th observation of the h th stratum in the sample

$$\sum N_h = N \text{ and } \sum W_h = 1$$

i. The Stratified sample mean

Given a population subdivided into H strata having unequal number of units, and noting that the interest here is in estimating the population mean and its variance, we first consider each of the stratum mean along with the corresponding variance. The properly weighted values (means and variances) can then be combined to give estimates of the population mean and variance [Cochran, 1977].

Consider the mean denoted by $\bar{y}_{ho}$, which is the mean of

the h th stratum, where the subscript 0 denotes that the selection is with simple random sample and $h = 1, 2, 3, \dots, H$.

$$\bar{y}_{h0} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \quad (2.16)$$

The combined mean, or the mean of any stratified random sample of elements is given by

$$\bar{y}_{st} = \sum_{h=1}^H W_h \bar{y}_{h0} = \sum_{h=1}^H W_h \cdot \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \quad (2.17)$$

Therefore, for Stratified random sampling the variance of the estimate $\bar{y}_{st}$ is

$$\text{Var}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^H N_h (N_h - n_h) \frac{S_h^2}{n_h} = \sum_{h=1}^H W_h^2 \frac{S_h^2}{n_h} (1 - f_h) \quad (2.18)$$

where, $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2$.

If the sampling fractions $\frac{n_h}{N_h}$ are negligible in the

strata, then the appropriate formula (when the finite population corrections are ignored) is

$$\text{Var}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^H \frac{N_h^2 S_h^2}{n_h} = \frac{\sum_{h=1}^H W_h^2 S_h^2}{n_h} \quad (2.19)$$

With proportional allocation the above Equation (2.19) reduces to

$$\text{Var}(\bar{y}_{st}) = \sum_{h=1}^H \frac{N_h}{N} \frac{S_h^2}{n} \left(\frac{N-n}{N} \right) = \frac{(1-f)}{n} \sum_{h=1}^H W_h S_h^2 \quad (2.20)$$

If a simple random sample of n_h elements is selected from the h th stratum, then

$$\text{var}(\bar{y}_{h0}) = (1-f_h) \frac{S_h^2}{n_h}$$

where,

$$S_h^2 = \frac{1}{n_h-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 = \frac{1}{n_h-1} \cdot \left[\sum_{i=1}^{n_h} y_{hi}^2 - \frac{\bar{y}_h^2}{n_h} \right] \quad (2.21)$$

Combining the variances of the stratum means, we obtain the variance of the sample mean $\bar{y}_{st}$, as

$$\text{var}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^H N_h (N_h - n_h) \frac{S_h^2}{n_h}, \quad (2.22)$$

and this is equivalent to

$$\text{var}(\bar{y}_{st}) = \sum_h W_h^2 (1-f_h) \cdot \frac{S_h^2}{n_h} \quad \text{where } W_h \text{ is the proper}$$

weight.

Hence the design effect is given by:

$$deff = \frac{var(\bar{y}_{st})}{var(\bar{y}_{srs})} ,$$

where $var(\bar{y}_{st})$ is the variance of the stratified sample mean

and $var(\bar{y}_{srs})$ is the variance of a simple mean under a simple

random sample of n elements from a population of size N as given in 2-1-3-a. The computed values of the standard errors and $deff(deft)$ values of nine continuous socioeconomic variables for the three areas are given in Tables 1a-2, 1b-2 and 1c-2.

ii. Difference of two Stratified sample means

Given two populations say P and P' each divided into different strata, we wish to estimate the variance of the difference of the means $\bar{y}_{st}$ and $\bar{y}_{st}'$ of continuous socioeconomic variables of the two populations P and P' , respectively.

Since the two samples are selected independently from their respective populations we have

$$Var(\bar{y}_{st} - \bar{y}_{st}') = Var(\bar{y}_{st}) + Var(\bar{y}_{st}') , \quad \text{where, } \bar{y}_{st} = \sum_{h=1}^H W_h \bar{y}_{ho} ,$$

$$\bar{y}_{st}' = \sum_{h=1}^{H'} W_h' \bar{y}_{ho}' , \quad Var(\bar{y}_{st}) \text{ is as given in Equation (2.22).}$$

A similar formula for $Var(\bar{y}_{st}')$ can be given.

Thus the sample estimate of the variance of the difference of the two means is given by

$$\begin{aligned} var(\bar{y}_{st} - \bar{y}_{st}') &= var(\bar{y}_{st}) + var(\bar{y}_{st}') \\ &= \sum_{h=1}^H W_h^2 (1-f_h) \frac{S_h^2}{n_h} - \sum_{h=1}^{H'} W_h'^2 (1-f_h') \frac{S_h'^2}{n_h'} \end{aligned} \quad (2.23)$$

Using the formulas for the sample estimates of the variances of the differences in Stratified and Simple random sampling the standard errors and deff(deft) values of five continuous variables for the three survey areas are computed and given in Table 3a-3.

2-2. Portability of standard errors

Portability refers to properties of the estimate that facilitate its use far from its source.

Standard errors computed for one statistic can be imputed directly only to essentially similar statistics based on similar subclass sizes from similar survey designs. They are specific to the estimate and depend on:

- the nature of the variables,
- their unit of measurement,
- the nature and design of statistics derived from the variables,
- size of the sample bases, which can vary greatly for subclasses,

- sizes of selections from sample clusters and
- nature and size of sampling units.

Design effects, are considerably more portable than standard errors. They are widely used to modify simple random estimates, say for the simple mean $\bar{y}$;

$se(\bar{y}_{srs})$ to guess some $se(\bar{y})$ as $(deft) \times se(\bar{y}_{srs})$. When

computing $deft(\bar{y}) = \frac{se(\bar{y})}{se(\bar{y}_{srs})}$, we remove the scaling effects

of the units of measurement and of the sample's aggregate size.

We sometimes assume that $deft_0 = deft_1$, where 0 and 1 denote different variables. This would usually serve better than would assuming either that $se_1 = se_0$ or that $se_1 = se_{srs}$. However, the assumption can also be misused, and the portability of defts has been naively overestimated by many. First, within the same survey many statistics are based on subclasses, whose sizes vary greatly. If $deft_0$ is computed from the entire sample, the $deft_1$ for the same variable on a subclass is often considerably smaller.

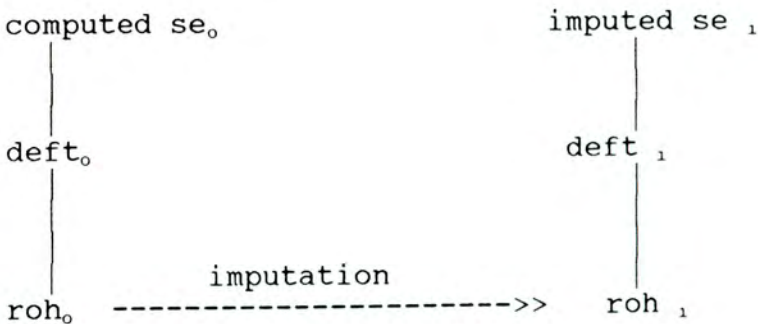
Design effects for most subclasses diminish along with sample size; and using values of $deft$ computed from the entire sample grossly exaggerates the actual effects of the design on subclasses. Second, $deft$ values depend heavily on the sizes of sample clusters used, which may differ drastically from one survey sample to another.

The design effect may be expressed as a function of roh

and $\bar{b}$ (the number of elements in sample clusters).

Since $deff = deft^2 = 1 + roh(\bar{y} - 1)$, as the size of the sample cluster increases (for large subclasses, or for new samples with larger cluster) for any variable, its $deft^2$ tends to increase also. We need portability to make inferences from one set of results to another set of variates with diverse values of $\bar{b}$ in the same survey or for designing another survey. Values of roh are more portable for this purpose than standard errors. The computed values of roh are functions of the kind of sampling units used and the nature of the selection process.

Thus, the proposed indirect method of imputation from a computed standard error (se_0) to an unknown one (se_1) according to Kish, Grooves and Krotki, (1976) is depicted as follows:



We impute across roh values because of their relative stability across diverse subclasses for each variable from a sample, and also for similar variables across samples.

3. ANALYSIS AND RESULTS

Table 1a-1

Mean, Standard error, Design effect and Roh values for continuous Demographic variables under Cluster Sampling Design, for RURAL area.

Variable	n	r	se.SRS	se.CL	deft	roh
Age	482	31.980	6.900	12.110	1.755	.026
No. of child. surviving	482	2.700	1.440	1.010	.701	-.007
No. of male children surviving	482	1.810	.800	.690	.863	-.004
No. of female children surviving	376	1.590	.860	.590	.686	-.008
No. of children in a large family	476	9.240	2.990	3.660	1.224	.006
No. of children in a small family	473	2.870	1.040	1.120	1.076	.002
Education level	482	.160	.590	.080	.135	-.013
No. of child. ever born	482	4.510	1.900	1.700	.894	-.013
Income of respondent	330	109.450	52.200	42.080	.806	-.006

The above results do not seem to exhibit any pattern on the efficiency of Cluster sampling vis-a-vis Simple random sampling. However, when dealing with continuous demographic data that have wide variation such as income, age, number of children ever born and surviving, Cluster sampling seems to be less efficient compared to Simple random sampling. Whenever there is little variability as is the case with educational level, female children surviving, then Cluster sampling seems to be more efficient than Simple random sampling. The overall result suggests that Cluster sampling is preferred over Simple random sampling when the data is categorized.

Table 1a-2

Mean, Standard error and Design effect values for continuous Demographic variables under Stratified Sampling Design, for RURAL area.

Variable	n	mean	se.SRS	se.ST	deft
Age	482	31.980	6.900	.320	.046
No. of child. surviving	482	2.700	1.440	.060	.042
No. of male children surviving	482	1.810	.800	.040	.050
No. of female children surviving	376	1.590	.860	.040	.046
No. of children in a large family	476	9.240	2.990	.140	.047
No. of children in a small family	473	2.870	1.040	.050	.048
Education level	482	.160	.590	.030	.051
No. of child. ever born	482	4.510	1.900	.090	.047
Income of respondent	330	109.450	52.200	2.840	.054

The defts in the above Table compare Stratified Random Sampling With Simple Random Sampling and clearly show the superiority of the former over the latter. Most interesting is the fact that there is no big variability in the value of the defts. They range between 0.042 and 0.054.

Mean, Standard error, Design effect and Roh values for continuous Demographic variables under Cluster Sampling Design, for GONDAR town.

Variable	n	r	se.SRS	se.CL	deft	roh
Age	509	34.280	8.950	14.760	1.649	.014
No. of child. surviving	512	3.540	1.879	1.533	.816	-.003
No. of male children surviving	512	2.190	1.082	.943	.872	-.002
No. of female children surviving	419	2.140	1.162	.922	.793	-.004
No. of children in a large family	500	9.270	3.040	3.961	1.303	-.003
No. of children in a small family	500	3.340	1.741	1.442	.829	-.003
Education level	512	2.170	1.700	.980	.576	-.005
No. of child. ever born	51	5.730	2.661	2.473	.930	-.001
Income of respondent	498	191.410	136.910	82.344	.601	-.005

Like the results of the Rural area, the results here show the superiority of Cluster sampling over Simple random sampling in 7 out of 9 cases, the ROH values are also much smaller. One should note that the Educational level of Gondar residents is higher than those of Rural area. There is less homogeneity among the sampled households in Rural area than the sampled households in the Gondar town. That is lower standard error among Rural residents than Gondar residents. The results suggest that taking larger sample or more clusters in Gondar town than in Rural areas is recommendable.

Table 1b-2

Mean, Standard error and Design effect values for continuous Demographic variables under Stratified Sampling Design, for GONDAR town.

Variable	n	mean	se.SRS	se.ST	deft
Age	509	34.280	8.950	.387	.043
No. of child. surviving	512	3.540	1.879	.081	.043
No. of male children surviving	512	2.190	1.082	.048	.044
No. of female children surviving	419	2.140	1.162	.055	.048
No. of children in a large family	500	9.270	3.040	.141	.047
No. of children in a small family	500	3.340	1.741	.072	.041
Education level	512	2.170	1.700	.067	.040
No. of child. ever born	512	5.730	2.561	.100	.038
Income of respondent	498	191.410	136.910	6.012	.044

When Stratified random sampling was compared with Simple random sampling in Gondar town, the superiority of the former is now lower compared with those in the Rural areas is shown by the lower value of defts (0.043 against 0.048).

Table 1c-1

Mean, Standard error, Design effect and Roh values for continuous Demographic variables under Cluster Sampling Design, for DABAT town.

Variable	n	r	se.SRS	se.CL	deft	roh
Age	212	31.450	6.840	15.096	2.207	.034
No. of child. surviving	212	2.940	1.649	1.411	.855	-.003
No. of male children surviving	212	1.950	.922	.933	1.012	.000
No. of female children surviving	140	1.890	1.131	.943	.834	-.004
No. of children in a large family	208	8.800	1.910	4.282	2.242	.039
No. of children in a small family	208	3.400	1.360	1.631	1.199	.004
Education level	212	1.630	1.559	.849	.544	-.007
No. of child. ever born	212	4.880	2.330	2.341	1.005	.000
Income of respondent	199	156.580	111.650	76.497	.685	-.005

Dabat is a smaller town compared to Gondar. When one looks at the defts the results for Dabat are similar to those of the Rural areas than those of Gondar. Like the Rural areas Cluster sampling seems to be more efficient whenever there is little variation in the observations.

Table 1c-2

Mean, Standard error and Design effect values for continuous Demographic variables under Stratified Sampling Design, for DABAT town.

Variable	n	mean	se.SRS	se.ST	deft
Age	212	31.450	6.840	.458	.067
No. of child. surviving	212	2.940	1.649	.100	.061
No. of male children surviving	212	1.950	.922	.063	.068
No. of female children surviving	140	1.890	1.131	.094	.084
No. of children in a large family	208	8.800	1.910	.141	.074
No. of children in a small family	208	3.400	1.360	.089	.065
Education level	212	1.630	1.559	.100	.064
No. of child. ever born	212	4.880	2.330	.173	.074
Income of respondent	199	156.580	111.650	7.912	.071

The superiority of Stratified random sampling is considerable, however such superiority is not as high as that of Gondar.

Table 2a-1

Mean, Standard error, Design effect and Roh values for discrete Demographic variables under Cluster Sampling Design, for RURAL area.

Variable	n	p	se.SRS	se.CL	deft	roh
Pregnancy status	464	.307	.021	.022	1.015	.000
Demand for additional children	446	.480	.024	.020	.843	-.004
Knowledge of birth control	317	.127	.019	.055	2.918	.145
Approval of birth control	438	.203	.019	.029	1.503	.018
Approval of abortion	441	.080	.013	.020	1.568	.020
Living standard compared to community	413	.183	.019	.040	2.119	.051
Living standard compared to family	409	.103	.015	.040	2.636	.086
Work other than household work	315	.622	.027	.040	1.463	.022
Ownership status of the house	415	.015	.006	.010	1.670	.026
Mean	406.444	.236	.018	.031	1.748	.041
Standard deviation	54.259	.200	.006	.014	.692	.048
Coefficient of variation	.133	.847	.333	.452	.396	1.171

All the above variables are a zero - one choice type of categorical data. On this result Cluster sampling is less efficient than Simple random sampling except for one variable, namely the desire for more children. The overall mean shows an inefficiency rate of 174.8 percent which is substantial. The results suggest that one needs to take larger samples if one uses Cluster sampling for the case where categorical data from rural areas are generated. In addition, the results suggest homogeneity of opinions in rural areas.

Table 2a-2

Mean, Standard error and Design effect values for discrete Demographic variables under Stratified Sampling Design, for RURAL area.

Variable	n	p	se.SRS	se.ST	deft
Pregnancy status	464	.307	.021	.021	.998
Demand for additional children	446	.480	.024	.024	.999
Knowledge of birth control	317	.127	.019	.017	.912
Approval of birth control	438	.203	.019	.019	.986
Approval of abortion	441	.080	.013	.013	1.012
Living standard compared to community	413	.183	.019	.019	.974
Living standard compared to family	409	.103	.015	.015	.980
Work other than household work	315	.622	.027	.027	.988
Ownership status of the house	415	.015	.006	.006	.971
Mean	406.444	.236	.018	.018	.980
Standard deviation	54.259	.200	.006	.006	.029
Coefficient of variation	.133	.847	.333	.333	.030

When Stratified random sampling is compared with Simple Random Sampling while collecting categorical data, there seems to be little difference.

Table 2b-1

Mean, Standard error, Design effect and Roh values for discrete Demographic variables under Cluster Sampling Design, for GONDAR town.

Variable	n	p	se.SRS	se.CL	deft	roh
Pregnancy status	494	.100	.014	.028	2.069	.027
Demand for additional children	465	.285	.021	.061	2.946	.067
Knowledge of birth control	470	.440	.023	.100	4.365	.155
Approval of birth control	489	.568	.022	.068	3.022	.067
Approval of abortion	493	.033	.008	.016	1.782	.024
Living standard compared community	477	.305	.021	.075	3.579	.100
Living standard compared family	471	.270	.020	.100	4.886	.196
Work other than household work	392	.550	.025	.058	2.289	.044
Ownership status of the house	456	.748	.020	.054	2.632	.052
Mean	467.444	.367	.019	.062	3.063	.081
Standard deviation	31.149	.231	.005	.028	1.044	.059
Coefficient of variation	.067	.629	.263	.452	.341	.728

As can be seen from the above table, when compared to the other areas Cluster sampling gives lower precision compared to Simple random sampling.

Table 2b-2

Mean, Standard error and Design effect values for discrete Demographic variables under Stratified Sampling Design, for GONDAR town.

Variable	n	p	se.SRS	se.ST	deft
Pregnancy status	494	.100	.014	.013	.989
Demand for additional children	465	.285	.021	.021	.982
Knowledge of birth control	470	.440	.023	.021	.897
Approval of birth control	489	.568	.022	.022	.966
Approval of abortion	493	.033	.008	.008	.980
Living standard compared community	477	.305	.021	.020	.965
Living standard compared family	471	.270	.020	.019	.908
Work other than household work	392	.550	.025	.025	.977
Ownership status of the house	456	.748	.020	.020	.968
Mean	467.444	.367	.019	.019	.959
Standard deviation	31.149	.231	.005	.005	.033
Coefficient of variation	.067	.629	.263	.263	.034

Like continuous variables the efficiency of Stratified sampling is slightly better than Simple random sampling. When this result is compared with the previous table (for cluster) the efficiency of Stratified sampling is much higher than Cluster sampling.

Table 2c-1

Mean, Standard error, Design effect and Roh values for discrete Demographic variables under Cluster Sampling Design, for DABAT town.

Variable	n	p	se.SRS	se.CL	deft	roh
Pregnancy status	207	.105	.021	.047	2.182	.037
Demand for additional children	207	.370	.034	.173	5.153	.249
Knowledge of birth control	174	.165	.028	.065	2.309	.050
Approval of birth control	205	.180	.027	.050	1.873	.025
Approval of abortion	205	.015	.009	7.517	.883	-.002
Living standard compared to community	208	.095	.020	.041	2.017	.030
Living standard compared to family	206	.195	.028	.055	1.995	.029
Work other than household work	178	.740	.033	.092	2.798	.078
Ownership status of the house	157	.490	.040	.100	2.500	.068
Mean	194.111	.262	.027	.904	2.412	.063
Standard deviation	19.186	.231	.009	2.480	1.156	.074
Coefficient of variation	.099	.882	.333	2.743	.479	1.175

Like Gondar and Rural areas the deft value of Dabat town are large, telling that Cluster sampling has greater variance. In this case other design would be preferred to estimate the population proportion.

Table 2c-2

Mean, Standard error and Design effect values for discrete Demographic variables under Stratified Sampling Design, for DABAT town.

Variable	n	p	se.SRS	se.ST	deft
Pregnancy status	207	.105	.021	.021	.995
Demand for additional children	207	.370	.034	.032	.956
Knowledge of birth control	174	.165	.028	.028	.961
Approval of birth control	205	.180	.027	.027	.994
Approval of abortion	205	.015	.009	.008	.989
Living standard compared to community	208	.095	.020	.020	.997
Living standard compared to family	206	.195	.028	.027	.991
Work other than household work	178	.740	.033	.032	.978
Ownership status of the house	157	.490	.040	.039	.975
Mean	194.111	.262	.027	.026	.982
Standard deviation	19.186	.231	.009	.009	.015
Coefficient of variation	.099	.882	.333	.346	.015

As that of Rural and Gondar areas, the deft values for Dabat under Stratified random sampling are slightly less than one. This shows that Stratified sampling to be the better design for estimation of population proportions. Note that, deft values for each of the three areas under Stratified random sampling for continuous demographic variables are very much smaller than those for discrete ones. This tells that Stratification shows very good results for Continuous variables than for discrete ones.

Standard error and Design effect values for the difference of means of five continuous variables in cluster and stratified sampling designs for the three survey areas are shown in Tables 3a-1, 3a-2 and 3a-3.

The following notations are used in the three Tables below:

$\bar{y}_R$ = mean of the variable for Rural area

$\bar{y}_G$ = mean of the variable for Gondar town

$\bar{y}_D$ = mean of the variable for Dabat town

$$se1 = se(\bar{y}_R - \bar{y}_G)$$

$$se2 = se(\bar{y}_R - \bar{y}_D)$$

$$se3 = se(\bar{y}_G - \bar{y}_D)$$

$deff_1$ = the design effect for the difference of means of
Rural area and Gondar town

$deff_2$ = the design effect for the difference of means of
Rural area and Dabat town

$deff_3$ = the design effect for the difference of means of
Gondar town and Dabat town

Table 3a-1

Simple random sampling

Variable	se1	se2	se3
No. of child. surviving	2.366	2.189	2.500
No. of male children surviving	1.345	1.221	1.421
No. of female children surviving	1.446	1.421	1.622
No. of child. ever born	3.270	3.001	3.537
Income of respondent	146.524	123.250	176.664

In the estimation of the mean differences of the three survey areas it can be seen in Table 3a-1 that the standard errors of the difference of the means of Rural and Dabat areas are smaller than the difference of the means of Rural and Gondar. The standard errors of the differences of means of Gondar and Dabat are very high.

Table 3a-2

Cluster sampling

Variable	se1	se2	se3	deft1	deft2	deft3
No. of child. surviving	1.838	1.738	2.083	0.775	0.794	0.831
No. of male children surviving	1.166	1.158	1.327	0.866	0.949	0.933
No. of female children surviving	1.095	1.114	1.319	0.755	0.781	0.812
No. of child. ever born	3.005	2.895	3.407	0.917	0.964	0.964
Income of respondent	92.472	87.305	112.394	0.632	0.707	0.632

The deft values for the difference of means are relatively small, showing that Cluster sampling is better than Simple random sampling.

Table 3a-3

Stratified sampling

Variable	se1	se2	se3	deft1	deft2	deft3
No. of child. surviving	0.116	0.118	0.129	0.049	0.054	0.052
No. of male children surviving	0.060	0.072	0.079	0.045	0.059	0.056
No. of female children surviving	0.071	0.104	0.110	0.049	0.073	0.068
No. of child. ever born	0.133	0.194	0.200	0.041	0.065	0.057
Income of respondent	6.649	8.407	9.937	0.045	0.068	0.056

Deft values for all of the three differences are small. This shows that Stratified random sampling is an appropriate design in the estimation of differences of two population means. In addition the results here show Stratified sampling to be superior than not only Simple random sampling but also Cluster sampling.

4 . CONCLUSION

The point of departure of practical sampling is the material and psychological constraints imposed by the data collection operation. These constraints strongly influence the choice of sample design. However, data relating to sampling errors (and to costs) can provide the designer with a more sensible basis for making specific choices among the options that are available.

These data are in any case of great value in interpreting the actual results of the survey and in evaluating the performance of the sample design.

In this paper an attempt was made to study the standard errors, design effects(defts) and the rate of homogeneity(roh) from the survey of the three areas Rural, Gondar and Dabat towns. Depending on the computed results of sampling errors in the survey, the following concluding remarks may give some insights into the degree of reliability and the homogeneity of responses as well as the effectiveness of Cluster and Stratified sampling designs over Simple random sampling in the socioeconomic surveys of the Rural settings of Ethiopia.

Almost all of the design effects(defts) obtained for stratified random sampling are less than 1.0 (Tables 1a-2, 1b-2 and 1c-2). They range from 0.042 to 0.054, 0.038 to 0.048 and 0.061 to 0.084 for Rural, Gondar and Dabat areas, respectively. Small values of defts occurred not only for continuous variables but also for discrete variables. In all of the cases Stratified random sample is more efficient than Simple random sampling. This is reassuring for the survey user.

When considering the continuous socioeconomic variables involved under Cluster sampling, the low values of deft and roh for most of the variables suggest that Cluster sampling is more efficient than Simple random sampling for Rural and Gondar areas (Tables 1a-1 and 1b-1). But for Dabat town the deft values are not much less than 1 for most of the variables (Table 1c-1). The roh values are small ranging from -0.007 to 0.039, this suggests that the sampled households in Dabat are relatively homogeneous even if the homogeneity is not as much as that of Gondar and Rural area in which above 63% of the variables receive homogeneous responses. This implies that Cluster sampling is relatively superior to Simple random sampling when the information deals with responses of variables which are of continuous type.

For categorical type of data (discrete variables) the deft values for Cluster sampling are much larger than that of continuous variables. Almost all of the deft values for the three survey areas exceed 1 (Tables 2a-1, 2b-1 and 2c-1). This in turn suggests that Simple random sampling is more efficient than Cluster sampling when dealing with responses of categorical type questions. With data based on the rural parts of Ethiopia high values of rohs and defts for such questions are to be expected; that is, due to different reasons it is not unusual to get some uniform responses.

In the estimation of the mean differences of the three survey areas the pattern of deft values of Cluster sampling showed relatively small values for Rural and Gondar areas, Table 3a-2. While Dabat has greater variability when the difference is taken from Rural and Gondar areas. In any case

the deft values for the difference of the means of Rural and Gondar area are less than 1, which implies that Cluster sampling is preferred in estimating the difference of two means.

Similarly the deft values for Stratified random sampling again confirm that the mean differences between Rural and Gondar have less variability than other differences namely the difference of Rural and Dabat, and that of Gondar and Dabat, Table 3a-3. In general it can be seen that all of deft values are less than 0.075, table 3a-3, which implies Stratified random sampling is more efficient in estimating the mean differences of the three areas for continuous socioeconomic variables than Simple random sampling.

Generally, we can conclude that the estimated sampling errors could be used for planning and designing of similar sample surveys making such studies scientific and precise. From the estimated deft and roh values in some cases we found that Cluster sampling is less precise implying that other designs would have been appropriate. Most of the estimated values indicate that Stratified sampling is the most efficient of the three designs. It seems that, from this result Cluster sampling would be an appropriate design when dealing with variables (questions) having responses of quantitative nature; and stratified sampling would be appropriate for both quantitative and qualitative types of variables.

References

1. Anderson, T.W., and Sclove, S.L. (1978), Introduction to the Statistical Analysis of Data, (Houghton Mifflin Comp, Boston)
2. Cochran, W.G. (1977), Sampling Techniques, 3rd ed, (Delhi, Wiley Eastern Limited).
3. Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), Sample Survey Methods and Theory, John Wiley and Sons, New York and London.
4. Kidane, A.(1984), Sampling errors from socioeconomic baseline survey in an Ethiopian rural setting, (Paper presented to the OSSREA Workshop on Research Methods in the Social Sciences in Addis Ababa, October 1-3, 1984).
5. Kish, L., Groves, R.M. and Krotki, K.P., (1976), Sampling Errors for Fertility Surveys, University of Michigan Ann Arbor, Michigan 48106.
6. Kish, L. (1965), Survey Sampling, John Wiley and Sons, New York and London.
7. McCarthy, P.J. (1957), Introduction to Statistical reasoning, McGraw-Hill Book Company, New York.
8. Som, R.K. (1973), A manual of Sampling techniques, Heiner, London.

9. Sukhatme, P.V. and Sukhatme, B.V. (1970), Sampling theory of Survey with applications, Iowa State University Press, U.S.A.
10. Verma, V., Scott, C. and O'Muircheartaigh, C. (1980), 'Sample designs and sampling errors for the World Fertility Survey', Journal of the Royal Statistical Society (A), Vol.143, pp. 431-473.
11. Yates, F. (1971), Sampling Methods for Censuses and Surveys, 3rd.ed., Griffen, London.
12. Zarkovich, S.S. (1965), Sampling Methods and Censuses, Food and Agriculture Organization of the United Nations, Rome.

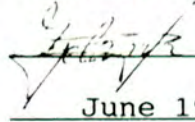
DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been submitted for a degree to any other university. All sources of materials used for the thesis have been duly acknowledged.

Name.

Zealealem Tarekegn

Signature.



Date of submission.

June 1994
