CHAPTER II

AN OVERVIEW OF COMMONLY USED SAMPLING DESIGNS

2.1 Introduction

Of all the aspects of sample surveys, sampling techniques are the most highly developed in terms of methodology and theoretical framework. Despite the relatively short history of sampling, there is extensive literature on the subject, ranging from abstract mathematical investigations to handbooks for the practical samplers. However, newcomers to surveys are sometimes puzzled as to how little the vast body of sampling theories is being applied in practice.

Survey sampling is partly science and partly an art. In ad-hoc surveys, the great majority of samples are designed with little or no reference to the elaborate methodology of optimization that has been worked out by the sampling theoreticians. But when a series of surveys are to be conducted regularly, the opportunity exists for applying sampling theory with a view to improving sampling designs on the basis of the experience gained in earlier surveys. However, in practice, analytical studies of variances, costs and organizational problems that are necessary for improving the efficiency of future surveys have rarely been undertaken.

This chapter reviews the important aspects in survey designs particularly in sampling designs. The purpose of this review is to delineate the sampling techniques to provide an analytical framework in the following chapters. In this review, essential factors that will impact the methods of analysis of the design effects of the different sampling schemes will be identified..

The review begins with an exploration of the main ingredients of sampling designs that are commonly used in nationwide surveys. Various selection techniques are reviewed with a brief discussion of the main features of each technique. This will be followed by reviews on the type of sampling frames used and type of errors in the sample surveys. Four common methods of variance estimation techniques for complex sample designs will be discussed. Lastly, the methods which are used to examine the sampling errors are reviewed.

2.2 Sampling Designs

The design of a survey involves many interrelated decisions on such factors as the mode of data collection (face-to-face interview, telephone interview or self-administration), the framing of the questions and the method of data processing, as well as the sampling design (Kalton, 1983).

The overall survey design includes the following important aspects.

i. The survey coverage

ii. The definition of the survey variables

iii. The method of measuring the variables to be studied

iv. The method of data collection

v. The method of analysis of survey data

vi. The utilization of survey results in decision-making.

vii. The selection of appropriate sampling design

viii. The budget and cost of data collection.

The sampling design basically consists of two aspects. In the first place, a decision has to be made on how to select sampling elements using appropriate sampling designs. To a large extent, the availability of sampling frame plays a

major role in determining the selection process. Second, one has to estimate sample statistics based on survey data in making inferences to the population values. Since sampling involves a fraction of elements from the population, statistics estimated from sample would invariably be subjected to certain degree of fluctuations. The average fluctuation, which is the standard error, can be estimated from the sample itself. Of course, apart from sampling fluctuation, the non-sampling errors also contribute to deviations from the true values of the population parameters. But the theory of sample surveys deals primarily with the standard error of estimates because statistical inference is based on such errors that can be estimated (Kish, 1965).

Various sampling designs concerning selection process and estimation of sample statistics and standard errors were presented in Yates (1971). His extensive discussions range from simple random sampling to multistage sampling including sampling with probability proportionate to size. Many examples were presented together with numeric calculation to facilitate the discussion of various sampling designs (Yates, 1971).

Among the designs, stratified multistage design is the most common type of designs used in large-scale surveys. This design combines the advantages of stratification and clustering when strata are made internally homogeneous and the first-stage units are internally heterogeneous, so that a small number of first-stage units need to be selected from each stratum in order to provide an efficient sample. Usually, the primary strata are compact geographical areas because it makes field enumeration easier. The ultimate units may be smaller homogeneous areas or groups of units having similar characteristics.

Multistage sampling will often give more efficient estimates than unrestricted simple random sampling because of the usual economics of clustering and better sampling plan for the second and subsequent stages. Also, the advantages of stratification are generally considerable and often out-weigh the slight increase in the complexity of analysis. The selection process can be designed according to the most efficient method applicable for each stage. Of course, attention should also be given to the advantages of adopting self-weighting design. The techniques of interpenetrating sub-samples may also be incorporated as an integral part of standard sampling designs. This technique enables one to:-

i. Examine the factor causing the variations.

ii. Compute the sampling error from the first-stage units if these comprise one level of inter-penetration.

iii. Provide advanced estimates on the basis of one or more sub-samples.

iv. Provide the basis of analytical studies by the method of fractile graphical analysis (Som, 1973).

2.3 Selection Processes

The simple random sampling is the simplest type of selection process and it is the basis of the more complex sampling methods. In stratified sampling, the population is divided into groups or strata. These strata may or may not contain the same number of units. The fraction of the units selected can be the same for all strata or vary by stratum depending on the gains in accuracy that will result if different sampling fractions are used for different strata. If the sampling units are not classified in the required strata, then the stratification process may be carried out after the selection.

Systematic selection is a much more practical sampling process even though the importance of the principle of random selection in sampling has been stressed. In systematic selection, it is assumed that a list of target population is available and then every k^{th} entry on the list will be selected. In multistage selection process, the first-stage sampling units are made up of a number of second-stage units. In every stage of the selection, one may use suitable sampling method together with appropriate sampling fractions. This method incorporates the flexibility of utilizing existing natural divisions and subdivisions of the sampling units at every stage in achieving the gains generally associated with stratified random sampling.

Another selection method is sampling with probability proportional to size of unit, which yields a self-weighting sample. This method is particularly attractive in agriculture survey in determining the proportion of points that fall in areas of target population. Multi-phase sampling is another selection strategy that allows one to collect certain information from parts of units in the sample and other information from sub-samples of the original sample.

In balanced sample method, the sample is selected in such a manner that a known quantitative character of the selected units is equal to this character of all the units of the population. This method is also known as purposive selection, which was at one time extensively used in sample surveys. As no rigorous rules of selection were followed, the sample may be unrepresentative in other ways. For this reason, purposive selection is not considered as a probability sampling and has largely been replaced by the principles of stratification in modern sampling.

A systematic area sampling is a selection process of land areas utilizing maps by superimposing a grid of points to form the sampling units. Line sampling is a method employed mainly in agriculture and forestry to select a sample from the

whole selected area block. The information can then be obtained for all points on the line (Yates, 1971).

According to Kish (1965), simple random sampling is the basic selection process and all other procedures can be viewed as modifications. Basically the modification procedure can be grouped into 5 major types with combination of them yielding a large variety of possible designs. This variety increases rapidly in multistage samples, which provides great flexibility in the choice of different kinds of frames and practical procedures.

The first type concerns the selection of population elements with equal probability (*epsem*) or unequal probability. Second type deals with the form of sampling units in the selection. In element sampling, the sampling unit contains only one element and the selection is performed directly on the elements. However, in cluster sampling, the sampling units are made up of a cluster of elements. This will result in multistage sampling as the selection of the elements follows the selection of sampling units in two or more stages. The third type concerns with partitioning of population into groups call strata. In unstratified selection, the sampling units are selected from the entire population. For stratified sampling, the population is divided into sub-populations and separate selection is performed within each stratum. The fourth type concerns the selection mechanism used to obtain the sample. A sample can be selected using random selection or systematic selection by selecting every k^{th} sampling unit. The last type focuses on the sources where the final sample is obtained. In one-phase sampling, the final sample is selected directly from the entire population. Alternately, in two-phase sampling the final sample is sub-selected from a pre-selected larger sample that provides information for improving the final selection. (Kish, 1965)

Of the modification procedures mentioned above, all the processes were outlined in Yates (1971), except in cases involving the combination of these procedures in a more complex designs. Kish (1965) had discussed in great details on multistage cluster sampling. In cluster sampling, each sampling unit contains more than one population elements. Each element must be uniquely identified to belong to only one sampling unit. Fortunately, there are many natural clusters, especially in the general public surveys where the ultimate sampling units of interest is an individual person. For example, family or dwelling unit can always be a convenient cluster of individual persons. The building block can be a cluster of dwelling units. Clustering should be preferred over element selection when the lower cost per element more than compensates for the higher element variance and greater difficulty in statistical analysis. This occurs often for large and widespread national samples. The choice of cluster is the recognition in the sampling design of some features in the physical distribution of the population and in the nature of the frame.

Multistage selection is employed in cluster sampling for a good compromise between the larger variance yielded by cluster sample and the precision gains in element sampling which always produce greater spread of samples over the population, taking cost into consideration. The aim in multistage selection is to reduce the degree of clustering which will directly decrease the variance and at the same time without incurring a proportional increase in cost. In multistage selection, the sampling units of the first stage are called primary sampling units (PSUs). The subsequent stages are called second-stage sampling units, third-stage sampling units and so on. The selection mechanism applied at every stage may vary depending on the availability of suitable of frame at every stage.

Cluster sampling designs are normally employed together with stratification ecause stratification has more advantages for cluster than for element sampling.)efinitely, the process of stratifying population by cluster is easier than for each ıdividual element. Secondly, the relative gains of proportionate stratification are ;reater for cluster than for element sampling from the same set of strata. Thirdly, it ; easier to select clusters without replacement within each stratum.

For many populations in the world, element sampling is not feasible due to ıigh travelling cost in covering widespread geographical areas, as well as the lack of ;ood listings. Therefore, cluster sampling is necessary in some forms particularly in he form of area sampling frames, which provide the best solution (Kish, 1965).

Among the selection processes, a combination of various designs are :ommonly used in large-scale national surveys. In practice, multistage samples are requently used because an up-to-date and accurate frame is often unavailable. Even vhen suitable sampling frames for the ultimate units are available, the multistage ıelection is preferred as the cost of field work in element sampling can be very high. 3esides, multistage selection also help in reducing response error and improving ;ampling efficiency by reducing the intra-class correlation coefficient (rho) in the ɔrimary sampling units. Stratified multistage selection is the most common ɔombination procedure in almost all national-wide sample surveys. Many examples were quoted in Som (1973) and Verma and Le (1996) from various countries.

The use of a combination of various sampling techniques can be illustrated by the following examples. The Family Expenditure Survey by the Ministry of Labour, Great Britain in 1961 had incorporated a multistage stratified design with a uniform overall sampling fraction (Kemsley, 1966). The Health Interview Survey, a survey program by the U. S. National Center for Health Statistics was conducted

ısing stratified multistage probability sample of persons (Bean, 1974). The)emographic and Health surveys conducted over 48 countries also adopted the tratified 2-stage selection with self-weighting design in most of the samples (Verma nd Le, 1996).

:.4 Sampling Frames

As alluded to earlier, the choice of a sampling design depends largely on the ıvailability of suitable sampling frames. Thus, a major and often expensive task in lesigning and implementing a national population survey is to establish and naintain a sampling frame containing the target population. A good frame would ist the population units such that it facilitates sample selection to achieve greater :fficiency (Bryant, 1974). Two obvious examples of frames are lists of households ır persons enumerated in a population censuses, and a map of areas of a country ihowing boundaries of area units (Som, 1973).

Appraisal of the available or obtainable frames must be made in choosing the ılternative sampling frames that are available. The Wold Fertility Survey (1975) locumented the following characteristics of a good sampling frame as follows:

. It should be exhaustive.

i. It should be non-repetitive.

ii. It should be up to date.

v. The units should be clearly and unambiguously demarcated.

v. The units should be traceable in the field.

vi. Besides, it is also desirable that the units should be fairly constant in size.

In practice, the ideal frame is seldom available and the survey sampler has to e on the lookout for imperfection. Kish (1965, pp53-59) provided a useful fourfold lassifications of potential frame problems and possible solutions.

Between frame of lists and frame of maps, areas mapping frames are superior o lists, no matter whether the list is of individuals, households or dwellings. Sampling frame of list of individuals or households is very hard to keep up to date. The movement of population and formation of households happens rapidly, causing he difficulty in locating the individuals or households. Even though the frame of ist of dwellings is considerably more durable than list of households or individuals, he list of dwellings may not cover recent development and the structural changes to existing dwellings such as from non-residential to dwelling use and vice versa.

On the other hand, area frame provides a more convenient and effective rame for cluster of dwellings and people. Mapping enables the area units to be clearly identified and traceable in field and these often serve as primary sampling nits. Different types of area frames may be available in different countries. The deal frame is a map on which small areas have been marked off using natural oundaries. In some countries, such maps have been made for administrative urpose and others may be the census enumeration blocks. In many countries such naps are available but the area units are likely to be relatively large and further mapping may be required to create suitably small area units (WFS, 1975). The use of area sampling frame can also be observed in other documentation. The Demographic and Health Surveys (DHS, 1987) indicated that area sampling of enumeration districts are the convenient frame available in most developing countries for the surveys. The technical study. of Household Income and

Expenditure Surveys also indicated that administrative units are the commonly available types of frame, which is often used for sample surveys (UN, 1989).

In area sampling, one has to use a sampling frame covering the smallest area units possible so that after the selection of the area units, one can proceed directly to the listing of households or dwellings. If the area units are too large, sub-selection of areas may be required. In this case, further mapping operation within the selected units is needed to divide them into smaller area units or ultimate area units. Then, a sample of the ultimate area units is selected before carrying out the listing process of households or dwellings within the selected areas (WFS, 1975).

How small an area unit can be an ultimate area unit so that it is more cost effective to initial listing of dwellings than to create smaller area units? In the World Fertility Survey, it was suggested that each ultimate area unit should consist 100 - 200 households (WFS, 1975). It was also proposed that if the area units are larger than this, further stage of area sampling should be introduced. An average of not less than 500 population (approximately 100 households) was proposed in Demographic and Health Survey especially in rural areas and in unplanned urban areas (DHS, 1987). In the Pakistan Demographic and Health survey, each enumeration block consists of approximately 200 to 250 households (PDHS, 1992).

2.5 Sampling Errors

In using estimates from sample surveys, one of the important questions asked is what is the precision of the estimator. In general, there are two sources of errors. The first type of error has to do with the measurement error, which is the errors arising from response errors or coding errors. The second type of error inherent in

urvey data is sampling errors, which are errors that attributed to sampling elements om a population instead of taking a complete census.

Sampling errors cannot be avoided, but they can be measured and controlled t tolerable level at known risk by probability designs. Thus, while a certain ercentage of samples of a given design will lead to wrong conclusions about the opulation, the analyst must rely on the low probability of occurrence of such berrant samples and act on the assumption that the selected sample is not one of hese (Semon, 1958 pp269).

The sampling errors of a survey will depend on the size of sample, the ariability of the study variables, the sampling procedure adopted and also the way n which the results are calculated. With a proper process of selection, the sampling rrors can be calculated from the detailed results obtained from the sample, based on he mathematical theory of statistical sampling.

An evaluation of the sampling errors would indicate the relative accuracy nd efficiency of the different sampling methods, which can be used to improve uture surveys. The development of these processes has changed sampling from a peculative and uncertain procedure to a method having definite and determinable recision (Yates, 1971).

Cochran (1977) noted that the estimation of the standard errors from a ample is used primarily for three purposes. First, it is used to compare the precision btained by simple random sampling with that given by other method of sampling. Secondly, it is used to estimate the size of the sample needed in future surveys. Lastly, it is used to estimate the precision of data collected from the survey. These points have also been given emphasis by UN (1989), stating that standard errors

ıould be part of the technical report for the survey. Kish (1965) suggested it is a ood survey practice to present standard errors together with survey results.

Non-sampling errors are also a very important source of errors in a survey. he importance of non-sampling errors has long been recognized and a great deal of esearch had been done on this aspect. Among some of the studies on non-sampling rrors may be mentioned that of Hansen, Hurwritz and Bershand (1961) and Bailer nd Dalenius (1970) (cited in Krewski, 1981).

The total survey design concept was discussed in Namboodiri (1978) with a eview on the development of this concept. This concept implies that a balanced llocation of the resources available for a given survey toward controlling the ıagnitude of each of the different error component sufficient to minimize the total rror of the estimate of interest (Namboodiri, 1978).

In fact, non-sampling errors are often larger than sampling errors especially or large samples. While admitting the importance of the non-sampling errors of urvey work, Kish (1996) acknowledged the difficulties of estimating such errors. Iowever, he emphasized that the sampling errors tend to predominate in small amples and also for the small subclass even of large samples (Kish, 1996).

2.6 Methods of Sampling Errors Estimation

The other aspect of sampling design concerns the estimation of sample tatistics and their variances. As the methods of estimation of sample statistics and heir variances depend on the selection process adopted, various estimation procedures were developed together with the selection process. Yates (1971) presented a whole range of estimation procedures both for sample statistics and their variances for the selection processes illustrated. The estimation methods of

stimators can be grouped into two categories. The first category uses the arithmetic ean of the sample values and second category is based on ratio method or gression method that using supplementary information on a quantitative character ates, 1971).

The computation of sampling errors needs to take into account the complex mpling design and the form of estimator employed. Yates (1971), Som (1973) d Cochran (1977) illustrated the used of ratio method in stratified selection, ratified multistage selection and selection with probability proportional to size, tilizing auxiliary information available for the designs that may increase the fficiency of estimators. But, the variance of the ratio method is complicated by the ct that its denominator is a random variable. As a result, the sampling variance of e estimate is only an approximation that is valid in large samples (Cochran, 1977 hapter 6; Kalton, 1983; Levy et al., 1991).

Several general approaches exist for estimating sampling errors of estimators ased on complex sampling designs. Basically, it can be grouped into two general lass of methods that have been developed expressly for this purpose, namely *inearization* and *replication* methods. The technique of linearization method was leveloped by Keyfitz, Woodruff (cited in Levy et al., 1991) and others based on the aylor series approximation namely Taylor Expansion or Taylor Deviation. On the ther hand, *replication* is a method where an estimate of the variance of an estimated opulation parameter is obtained by expressing the estimated population parameter s a sum or mean of several statistics of subset sample observations. The procedures vailable are independent replication, pseudo-replications and Jackknife method Levy et al., 1991).

5.1 Taylor Expansion

Suppose that x_i has mean θ_i, and that the variances and covariances of the k riate ($x_1, x_2, \ldots, x_k$) are of order n^{-r}, $r > 0$. Consider the function $g(x_1, x_2, \ldots,$), which we write $g(x)$ for brevity.

$g'_i(\theta)$ is $\partial g(x) / \partial x_i$ evaluated at $(\theta_1, \theta_2, \ldots, \theta_k)$, we have the Taylor expansion as low:

$$(x) = g(\theta) + \sum_{i=1}^{k} g'_i(\theta)(x_i - \theta_i) + O(n^{-r}) \ldots (1)$$

ius, since $E(x_i) = \theta_i$,

$$\{g(x)\} = g(\theta) + O(n^{-r}).$$

not all the $g'_i(\theta) = 0$, we have further as below:

$$\text{ar}\{g(x)\} = E\left\{\left[\sum_{i=1}^{k} g'_i(\theta)(x_i - \theta_i)\right]^2\right\} + o(n^{-r})$$

$$= \sum_{i=1}^{k} \{g'_i(\theta)\}^2 \text{var}(x_i) + \sum \sum_{i \neq j=1}^{k} g'_i(\theta) g'_j(\theta) \text{cov}(x_i, x_j) + o(n^{-r}) \ldots (2)$$

Given that g is a ratio of two random variables where $r = (y / x)$ with the ndition of $x > 0$ if it is discrete and $x \geq 0$ if it is continuous (Kendall et al., 1958 hapter 11 for the condition). Then the equation (2) becomes

$$\text{ar}(y / x) = \frac{\text{var}(y)}{\theta_x^2} + \frac{\theta_y^2 Var(x)}{\theta_x^4} - \frac{2\theta_y \text{cov}(y, x)}{\theta_x^3} \ldots (3)$$

$$= \left\{\frac{E(y)}{E(x)}\right\}^2 \left\{\frac{\text{var}(y)}{E^2(y)} + \frac{\text{var}(x)}{E^2(x)} - \frac{2\,\text{cov}(y, x)}{E(y)E(x)}\right\} \ldots (4)$$

/here $E(y) = \theta_y$ and $E(x) = \theta_x$.

Equation (4) is the sum of the squares of the coefficients of variation of the /o variables, minus twice the square of what may analogously be called their ıefficient of covariation (Kendall et al., 1958).

The method will be used to obtain a variance estimator for the combined tio estimator (r), an important estimator of sample survey utilizing stratified ultistage design with notation as follow:

et $r = \bar{y}/\bar{x} = \Sigma y_h/\Sigma x_h = \Sigma\Sigma y_{hj}/\Sigma\Sigma x_{hj}$,

'here $E(y) = fY$, $E(y_h) = fY_h$, $E(a_h, y_{hj}) = fY_h$,

nd $E(x) = fX$, $E(x_h) = fX_h$, $E(a_h, x_{hj}) = fX_h$ with f is a constant and a_h primary ılections in stratum h.

The variable x is often the sample size or the sum of the weights, in which ase r is a ratio of mean or percentage.

et $r = g(y,x) = y/x$, its linear substitute is $g_l = (y-Rx)/(fX)$, where $R = Y/X$ and its pproximate variance is $V(r) = \{V(y) + R^2V(x) - 2Rcov(y,x)\} / (fX)^2$. On substituting ımple estimators for the unknown parameters in $V(r)$, a variance estimator is btained as equation below:

$$v(r) = \frac{\left\{v(y) + r^2 v(x) - 2r\,\mathrm{cov}(y,x)\right\}}{\bar{x}^2} \quad \ldots(5)$$

/ith appropriate estimates of $v(y)$, $v(x)$ and $cov(y,x)$ computed from the sample.

Woodruff (1971) had simplified the procedure by expressing $g(y,x)$ in terms ıf weighted values for the primary selections. Thus, at some point of forming the 'ariance estimator the sample values r and x can be substituted for R and fX. This ubstitution gives $g'_l = \Sigma\Sigma(y_{hj} - rx_{hj})/x$ where now the approximate linear substitute is

pressed as a straightforward sum, involving only the calculation of a summary of riable and a single variance (cited in Kalton, 1977)

Thus, this procedure is used to replace the calculation of $v(y)$, $v(x)$ and $v(y,x)$. The computation technique of calculating $z_h = y_{hj} - rx_{hj}$ is widely used for mplicated statistics. Then, the variance in equation (5) can be simplified as llow:

$$\text{ır}(r) = \frac{1}{x^2}\sum_{h=1}^{H}\frac{1-f_h}{a_h-1}\left(a_h\sum_{j=1}^{a_h}z_{hj}^2 - z_h^2\right) \quad \ldots (6)$$

Besides providing variance estimators for mean and percentages based on the tal sample, the above equation can also be used for subclass ratio means by noring sample elements not in the subclass (Kalton, 1977; Kish, 1965 Chapter 6). ıis method was used in World Fertility Survey and the Demographic and Health ırveys (Kalton, 1977; PDHS, 1992; Verma, et al., 1996).

6.2 Independent Replications

Under this method, it is assumed that several independent samples are ɔtained from the same population with the same design. Thus, the overall sample designed as the combination of a set of r independent sub-samples or replicates, ıd independent estimates of the same statistic x_i can be calculated for each ɔplicate. This method is known as interpenetrating sampling was introduced by lahalonobis (1944, 1946) to facilitate the analysis of sampling as well as non-ımpling errors. Deming (1956, 1960) advocated the use of replicated sampling ıethods for the ease of sampling error computations (cited in Kalton, 1977).

Suppose that ($x_1, x_2, \ldots, x_r$) are statistics for each replicate, then the mean stimate and variance of estimate would be as below:

Mean, $\bar{x} = \sum_{i=1}^{r} (x_i / r)$ and

Variance, $\text{var}(\bar{x}) = \sum_{i=1}^{r} (x_i - \bar{x})^2 / r(r-1)$ ... (7)

The attraction of replicated sampling for variance estimation is that the riance for $\bar{x}$ can be estimated using the above variance formulation regardless of e complexity of the sampling design within the replicates and the complicated rm of the estimator. Besides, if the estimates of the separate replicates are esented together with the overall estimate, it demonstrates the sampling variability that estimate.

However, in practice, the estimator to be computed is $\hat{x}$ which is the timator obtained using data from all the sub-samples, and its variance is $\text{var}(\hat{x})$. general, especially for complex statistics, $\hat{x}$ will not be equal to $\bar{x}$, but it is cessary to assume that $\text{var}(\bar{x})$ is approximately equal to $\text{var}(\hat{x})$. Of course, if $(x_1, x_2, \ldots, x_r)$ are linear estimator, then the overall estimator $\hat{x}$ obtained by pooling e replicates is equivalent to $\bar{x}$. In applying this method, $\hat{x}$ is always being used the estimator with its variance being approximated by $\text{var}(\bar{x})$.

The practical problem with this technique is that it may result in a less ficient sampling design by restricting the amount of stratification that can be nployed. This is because each independent sample is much smaller than the verall samples and this restriction is particularly serious for multistage designs. On ie other hand, a conflicting concerns with above problem is the choice of the umber of replicates, r. Definitely, larger number of replicates is preferred because will provide larger number of degree of freedom for the estimate of the variance nd improve its precision. But, it is unlikely to have large number of replicates with

ge sample size for each replicate. In fact, the number of independent replicates is ely to be small for a nationwide population survey (Kalton, 1977).

.3 Pseudo-replications

This method is also known as balanced repeated replication (BRR) in most the textbooks. This method is designed with exactly two PSUs per stratum in the mple. A random half of the sample is defined by randomly selecting one of the Us in each stratum and assuming the half sample and its complement are oroximately independent samples.

Suppose that a variance estimator for an estimator of the population total is be obtained using this technique. Let $2y_{hj}$ be the estimator of population total of atum h from primary selection j. Let $y^1 = 2\sum y_{h1}$ and $y^2 = 2\sum y_{h2}$ be the imators of Y from the selected half sample and its complement.

en the estimator, $y = \frac{1}{2}(y^1 + y^2)$ and its variance is

$$r(y) = \frac{1}{2}\left\{(y^1 - y)^2 + (y^2 - y)^2\right\} = (y^1 - y)^2 \ldots (8)$$

As the above estimator is only having one degree of freedom, the operation drawing half samples can be repeated T times to obtain a more stable estimator th the average of the T variance estimates computed as:

$$\text{var}_T(y) = \sum (y_t^1 - y)^2 / T$$

tting $\alpha_{th} = 1$ if j(t) = 1 and $\alpha_{th} = -1$ if j(t) = 2, to determine the particular half mple that is used. Then, $Var_T(y)$ can be expressed as below (Kalton, 1977):

$$ar_T(y) = \sum (y_{h1} - y_{h2})^2 + \frac{2}{T}\sum_t \sum_{h<k} \sum \alpha_{th}\alpha_{tk}(y_{h1} - y_{h2})(y_{k1} - y_{k2}) \ldots (9)$$

McCarthy (1966) suggested that pseudo-replication method can be employed hree different scenarios. First, the half sample replication based on formula (9) is :d to calculate the variance of the estimator. Second, using the balanced half nple replication such that if the set of half samples is selected in such a way to isfy the condition $\sum_{l} \alpha_{lh}\alpha_{lk} = 0$ for all pairs *(h, k)*, the second term on the right of iation (9) will disappear. Lastly, the partially balanced sample using a subset of tially balanced half samples to produce the variance of estimates, such as iding the strata into two halves and balancing the replicates within but not ween the two halves.

Of the pseudo-replication approaches outlined above, the balanced repeated lication (BRR) is more widely used in practice. This can be evident that many tbooks such as Cochran (1977), Kalton (1983) and Levy et al. (1991) focus the cussions on BRR method. In Kalton (1983), a selection of references on the hnique is provided. This method is used by National Center for Health Statistics CHS) in the Health Examination Survey (McCarthy, 1966; Bryant, 1974).

5.4 Jackknife Method

The 'jackknife' approach proposed by Quenouille (1956) and Tukey (1958) is intuitive approach in computing variances (cited in Namboodiri, 1978). This ethod involves splitting of total sample into a set of sub-samples and then opping out each of the sub-sample in turn.

et $\bar{z}$ be the estimator of the population parameter from the total sample.

et z_k' denote the estimator form the k^{th} jackknife replicate and z_k'' the complement. hen, the variance estimators in different form are as below:

$$var_1(\bar{z}) = \frac{1}{2}\sum\left\{\left(z_k' - \bar{z}\right)^2 + \left(z_k'' - \bar{z}\right)^2\right.$$

$$var_2(\bar{z}) = \frac{1}{2}\sum\left(z_k' - \bar{z}\right)^2$$

$$var_3(\bar{z}) = \frac{1}{2}\sum\left(z_k'' - \bar{z}\right)^2$$

$$var_4(\bar{z}) = \frac{1}{4}\sum\left(z_k' - z_k''\right)^2$$

These four variances are identical for linear estimator. For non-linear imators, an appraisal has been done by Frankel (1971) and Kish and Frankel)74) (cited in Cochran, 1977) using Current Population Survey from U. S. Census reau. The result indicated that $var_1(\bar{z})$ is the best of the four forms of variance imation and the same also for BRR method.

In summary, the Independent Replication method has the limitation to obtain r number of replicates to give reasonable precise variance estimators. The BBR :thod requires design with a paired selection and relatively small number of strata, ıile in Jackknife method excessive computation is required in calculating the riances. After taking into account the computing economy, applicability for the ıuired statistics and the sampling design employed, the Taylor Expansion method ll be used in this paper to compute the sampling errors for Media Index Survey.

7 Design effects and Intra-class Correlation

Examination of sampling errors should also be an important task in sampling :sign. There are always too many variables in the surveys that make it impossible compute the standard error for every single variable in the surveys. Therefore, odeling of survey sampling errors can be used to provide sampling errors for other atistics. Besides, the models may also help in planning of future sample design.

One approach is to plot survey estimates against some function of their ıpling errors and to fit a smooth curve to the resulting points. A well known thod is to compute relvariance, R_i and then examine the relationship between R_i l x_i where x_i is the sub-population total (Kaiton, 1977).

Another approach is to examine the relationship between the variance of a tistics of complex design with its variance from SRS sample. This ratio measures : design effect of the sampling design adopted. Kish (1965) defined the design ect (*deff*) as the ratio of the actual sampling variance, taking into account the mplexity of the sampling design to the variance of the same size (n) under iumptions of simple random sampling. Often, its square root, $deft = \sqrt{deff}$, iich is the ratio of standard errors is used.

$$ft^2 = \frac{complex\ \text{var}\ iance}{SRS\ \text{var}\ iance} = 1 + roh(b-1) \ldots (10)$$

The coefficient of intra-class correlation (*roh*) is the measures of mogeneity of elements of the primary selections and b is the number of sample ements per cluster. In unequal sized cluster, the average size $\bar{b} = n/a$ (a is the mber of cluster sample) also generally yields serviceable approximation. The ghest possible value of *roh* is +1. This corresponds to complete homogeneity of ements within clusters. On the other hand, the lowest possible value for roh is {-(b-1)} (Kish, 1965).

Kish et al. (1976) also proposed the use of indirect method of imputation om a computed standard errors to an unknown one. The imputation is done rough *roh* values because of the relative stability across samples and subclasses.