Chapter 1

INTRODUCTION
Chapter 1

INTRODUCTION

1.1.0 INTRODUCTION AND REVIEW OF LITERATURE

Sample surveys are performed to gather a variety of information about several unknown characters of the population. In areas of business, agriculture, marketing, labour forces etc., surveys are extremely popular and a first hand source of information for possible prediction. As per theory and experiences, surveys are faster than the complete enumeration of the population and also require lesser cost. In addition, these involve well trained man-power for data handling in order to produce greater accuracy in results due to least incorporation of non-sampling errors. The sampling error reduces with the increase of sample size but, at the same time, this increment enhances the amount of non-sampling errors. The theory of sampling is being developed as a mathematical formulation of estimation procedures for producing likely to accurate prediction under the prefixed cost when appropriate sample size and sampling procedure is used. Accuracy of the sampling

1
procedure is examining by the frequency distribution generated for the estimate if the procedure is applied repeatedly. During the mathematical analysis and discussion, it is customary to assume the target population with a finite size and then to look for the development of estimation techniques for the population parameters like mean, variance, coefficient of variation etc.

A sample may be drawn by judgment, by a random procedure or by a mixture of these two. The scientific basis could be provided only to the procedure of random sample and the most fundamental is called simple random sampling (SRS) which is affected a lot by the uncertainty of proper representativeness from population. The randomization opens avenue for the application of theory of probability among units in the sample selection and analysis procedure.

Cochran et.al. (1954) defined a probability sampling scheme where (i) every unit in the sample has a pre-assigned probability of inclusion (ii) the sample selected by a process involves one or more steps of automatic randomization consistent with these probabilities. The expectation with the sample user is to predict approximately the amount of error involved in the estimate derived from the gathered sample. The combination of probability theory and statistical inference has made possible to explore techniques for drawing valid conclusions and error estimation based on the sample information.
1.2.0 DATA COLLECTION

For any government, semi government or a private organization, the collection of data is a tool for looking into hidden facts and figures about the target population. By virtue of sample surveys, with structured or unstructured questionnaires, investigators come across a huge set of information, some may be of main interest and others be for additional interest. Also, survey deals with a variety of data, different methods of data collection, analysis and ways of interpretations. A most suitable selection among them is the objective of survey planning in order to predict with maximum amount of certainty. One can visualize a survey like a scientific experiment which needs proper planning and execution. The data may be classified into two categories (i) “survey data”, comprising of data which are collected and recorded by observation or inquiry, and (ii) “experimental data” which can only be obtained through a well designed and controlled statistical experiments.

1.3.0 SAMPLING TECHNIQUES

The population possess characteristic of heterogeneity, similarity and containing extreme values, all these affect the estimation procedures up to a great extant. While almost every observation in the population is nearly same, the smallest ever sample could produce a precise estimate whereas with large observational variations in the population a sufficiently large size sample may not be precise enough. This motivates
for creating a variety of sample schemes for different situations and kinds of data of the populations. A particular method of drawing sample may be good for one type but may not for other. The cost factor and time limitation are also vital aspects in this connection alongwith pre-fixation of sample size. So, the sampling technique has objectives (i) suitable methods of selecting a sample form the population, termed as "sampling procedure" and (ii) designing methods of computation about the desired characteristics of population from the sample termed as "estimation procedure".

The judgment sampling is completely affected by the biasness of individuals but the random sampling has a scientific basis like involving the probabilistic determination of procedure of unit selection and estimation strategies so as to infer about the unknown characteristics of the population bearing the least amount of errors. Generally, survey practitioners are concerned with the development of better sampling systems incorporating both, the survey design and estimation strategy. A sampling system could be more efficient than other if variance or mean squared error of the former is lesser to the later subject to the condition of the cost constraint.

The setup of basic design, simple random sampling is divided into two components based on the way of unit selection. If the unit is replaced in the population before the next draw, it is "simple random
sampling with replacement (SRSWR)" scheme. While selected units is not replaced before the next draw, it is "simple random sampling without replacement (SRSWOR)" and observed mathematically more efficient than SRSWR provided the sample size is greater than unity.

When population contains heterogeneity among units in terms of values, survey users are advised to form several homogenous groups and the design is well known called stratified sampling. All designs, other than these, are generated as a further modification of simple random sampling and stratified sampling. Some of important and well known are systematic sampling, cluster sampling, multiphase sampling, multistage sampling, area sampling, quota sampling, successive sampling, interpenetrating sub-sampling etc. (Kish (1967)). Apart from equal probability schemes where each unit has equal probability of selection, some times, situation demands for unequal probability selection for different elementary units. The consequence produces a series of varying probability plans and reveals that higher precision may be achieved by making probabilities of selection unequal. When units vary in size and the variable under study is prima facie correlated with size, probabilities of selection may be assigned in proportion to the size of the unit and scheme is termed as "probability proportional to size (PPS)" sampling. However, the theory of varying probability sampling is consequently more complex than that of equal probability sampling.
1.4.0 CLUSTER SAMPLING

A cluster implies a bunch of elementary units and in most of cases, units in clusters are geographically grouped at one place, reducing the cost of travel during data collection. A simple random scheme may contains units geographically scattered around the whole population which is contrary. The cluster sampling advocates for choosing bunch of units together to reduce the component of expenditure likely to occur during travel in surveys. One can assume a population containing several clusters (i) each of equal size or (ii) each of unequal size. In case of unequal clusters, the combining weight may be decided either by the cluster size or by the strata size in the set-up of stratified population.

1.5.0 ASPECT OF POST-STRATIFICATION

For a heterogeneous population, a simple random sample does not ensure to be a good representative, therefore, survey practitioners suggest for the stratification. The mathematical finding supports the fact that stratified sampling is more precise than SRS and often in administrative work the natural strata are found like villages, districts, zones, head-quarters etc. In most of stratifications, the population is grouped on the basis of additional information may or may not be related to the information of the main interest.
While the practical applications of stratified design needs the knowledge of (i) strata size and (ii) strata frames. If size is not available or frame otherwise, the implementation of stratified sampling is not in way. When stratum size are known but not the frames, the post-stratification is a solution and according to Sukhatme et.al. (1984) the post-stratification is as precise as the stratified sampling with proportional allocation when sample size is sufficiently large. According to Jagers et.al. (1985) the post-stratification with respect to relevant criteria may improve estimation subsequently over sample mean or ratio estimator. In light of these one could expect the post-stratification as a design closer to the real life situation.

The post-stratification could be mixed with usual cluster sampling design (See Ghosh (1963)) while assuming strata containing clusters as units and a random sample of clusters is chosen for the post-stratification. It constitutes post-stratified cluster design to be utilized for the estimation purpose.

1.6.0 REVIEW OF LITERATURE

It was some outstanding contributions of Mahalanobis, Neyman and Sukhatme during the thirties which brought the sampling theory in to academic horizon. Their sincere efforts have opened up several new avenues in the philosophy and settled foundations of the topic. With the introduction of use of auxiliary information in estimation strategies by
Cochran (1942), Henson and Hurwitz (1943) researches in sampling twisted to a new shape during the forties. Madow and Madow (1944) developed the technique and procedure of systematic sampling. During 1950’s, some aspects were unfolded by Horvitz and Thompson (1952), Godambe (1955,60), Koop (1963), Basu (1958) etc. which generated so-called ‘unified theory of sampling’ [Review by Godambe (1965)]. Contributions after 1960 are spread into wider dimensions and have the incorporation of model approach and Bayesian approach for improving the efficiency of various estimation procedures.


(A) BASED ON CLUSTER SAMPLING, STRATIFICATION, DEEP STRATIFICATION AND POST-STRATIFICATION

The stratification under the multi stage design and the involvement of prior information was considered by Ericson (1965) and Reddy (1976). Sethi (1963) and Ajagoankar (1971) studied the problem.
of searching the optimum points in the setup of stratification. The optimum allocation of the sample size using some prior distribution is taken up by Rao (1977).

The post-stratification is brought by Holt and Smith (1979) and further discussed by Jagers, Oden and Trulson (1985) and Jagers (1986). They have examined the utility of ratio estimator in post-stratification with the critical comment on bias caused. Smith (1991) made a useful contribution bringing a general note on post-stratification technique. Agrawal and Panda (1993) proposed a new combining weight structure for clubbing different post-stratified strata means. Agrawal and Panda (1995) elaborated the earlier estimation procedure in post-stratification with another approach. The cluster sampling is an effective design whose efficiency is examined by Singh (1956) whereas Ghosh (1963) considered the effect of post-clustering when the sample is by SRS. Sendransk and Akar (1979) picked up the estimation problem under post-stratified cluster sampling design and in similar approach a vital contribution is by Mehrotra, Srivastava and Dagi (1984).

The stratification converts into deep-stratification when every strata has more than one further sub-divisions (stratifications). In particular, with only one sub-divisions the stratification is called a two-way deeply stratified set-up considered by Bryant (1955). Further, in a useful contribution Bryant, Hartley and Jessen (1960) studied the estimation problem of population parameters under two-way stratified population.

(B) AUTHORS’ CONTRIBUTION IN SAMPLE SURVEYS

Shukla and Dubey (2001) explored a PSNR sampling scheme to handle the partial non-response and for tackling the complete non-response in mail surveys. A similar useful contribution is by Shukla and Dubey (2000). Shukla and Trivedi (1999) discussed post-stratification based on varying sample proportion in the weight structure. Shukla and Trivedi (2000, 2001) proposed post-stratification scheme for the mean estimation in a two-way stratified population. A one-parameter family of factor type estimator was suggested by Singh and Shukla (1987) and its further extension to the case of negative correlation discussed by Shukla, Singh and Singh (1991) and Singh and Shukla (1993). The study of factor type estimator in two phase sampling is noticed due to Shukla (2001) and factor type estimator with chaining technique is due to Singh, Singh and Shukla (1994). The factor type product estimator is discussed by Singh and Shukla (1995) and its extension to

(C) SOME OTHER INTERESTING CONTRIBUTIONS IN SAMPLE SURVEYS


1.7.0 PROBLEMS DISCUSSED IN SUBSEQUENT

CHAPTERS

The contribution in the thesis is focused around the estimation of population mean under the various setups of sampling designs using post-stratification. Sukhatme et. al. (1984) and Jagers et.al. (1985) strongly advocated for the wider applicability of post-stratification in situations where the strata frames are not in hand.

Bryant (1955) and then Bryant, Hartley and Jessen (1960) have initiated the problem of mean estimation under two-way setup which needs to be extended under post-stratification. Agrawal and Panda (1993, 95) suggested a new weight structure to combine different strata means for constructing more efficient estimator than the usual in post-stratification.

The Chapter 2 deals with the problem of mean estimation in post-stratification subject to condition of availability of prior information. It has a newly proposed grouping strategy as a solution where the estimator fails to provide efficient result. A class of estimators is proposed with a weight structure $C_iW_i$ to be chosen subject to the
condition of minimum mean squared error. A concept of PPM matrix is introduced and its properties, applications are derived. Necessary condition for the gain over usual estimator is obtained and cases where this does not satisfy, a general grouping strategy is proposed.

The Chapter 3 incorporates the stability property of the ratio $M_1/M_2$ where $M_1$ and $M_2$ are defined under usual notations in stratified sampling setup as

$$M_1 = \left( \frac{1}{N} \right) \left[ S^2 - \sum_{i=1}^{t} W_i S_i^2 \right]$$

$$M_2 = \frac{1}{(N - 1)} \left[ \sum_{i=1}^{t} (1 - W_i) S_i^2 \right]$$

It is assumed a survey is repeated over several occasions under the same object of estimation in order to learn the rate of change over time. The ratio $M_1/M_2$ is found having tendency of stability among occasions if time span is not very large. This could be guessed by past experience, pilot survey or otherwise due to being a stable quantity and could be utilized for the purpose of efficient estimation. The content of the chapter shows an application of this property.

In the Chapter 4, a two-way deeply stratified population of the classification type $r \times r$ and $r \times s$ is considered. The assumption is that size of strata as well as frame, both are unknown and the only known information is about the proportion of row and column-size-totals of the
two-way deep-stratification. A biased estimator is proposed and its properties are examined. Relative comparisons of efficiencies are performed. It is concluded that fifty percent of the sum of row-size-total proportions and the same of the column-size-total proportions provides an easy choice of unknown constant.

Next Chapter 5 contains the estimation of population mean in the set-up of post-stratified cluster sampling. The chapter contains a comparative study of several post-stratified cluster estimators and a class of those estimators with optimum properties.

The Last Chapter 6 also contains the estimation of population mean in the of post-stratified cluster sampling but using a different weight structure for combining cluster means. A class of estimator is proposed and a comparative study with other estimators is performed. Derived results in the mathematical form are numerically supported.

The five appendices A, B, C, D and E enclosed containing some relevant documents in support of significance of the research contribution of the thesis.