V84
U95

SOME STUDIES ON NON-OVERLAPPING CLUSTERS OF TWO UNITS

H.K. SHARMA

...ral Research Statistics New Delhi-12
ACC N. R.2319.
043:286
Library

Dissertation submitted in fulfilment of the
requirements for the award of Diploma
in Agricultural Statistics of the
Institute of Agricultural
Research Statistics
New Delhi

DE-04
9

INSTITUTE OF AGRICULTURAL RESEARCH STATISTICS
(Indian Council of Agricultural Research)
NEW DELHI-110012
1 9 7 4

ACKNOWLEDGEMENTS

I have great pleasure in expressing my deep sense of gratitude to Shri S.K. Raheja, Senior Statistician, Institute of Agricultural Research Statistics, New Delhi, for his valuable guidance, keen interest and constant encouragement throughout the course of investigation and preparation of the thesis.

I am grateful to Dr. B.B.P.S. Goel, Statistician-cum-Associate Professor, I.A.R.S., New Delhi, for his valuable suggestions in the course of study.

My thanks are also due to Dr. D. Singh, Director, Institute of Agricultural Research Statistics, New Delhi, for the facilities provided for carrying out the investigation.

The financial assistance by way of fellowship provided by the I.C.A.R. is gratefully acknowledged.

Hari Kant Sharma

(H.K. SHARMA)

CONTENTS

CHAPTER - I.

INTRODUCTION

1.1 For formulating an efficient sampling design not only the choice of appropriate strata and the procedure of selection are important but also the choice of a convenient sampling unit on which charactor values are to be measured. The sampling unit is defined as an element or group of elements which is identifiable and conveniently observable. Quite often it is not practicable to measure the characteristic under study on the elements of the population sometimes it may be profitable to consider a group of adjacent units as a sampling unit even though each unit of this group may itself be an identifiable observational unit and hence may as well be considered as a sampling unit. In actual practice, however, generally no reliable list of the elements in the population is available and it is very difficult and expensive to construct such a list. Particularly when the population is very large, we cannot identify the units easily and in that case they have to be identified by means of groups of units or larger units. For instance, in a city a list of all the houses is readily available but that of persons is rarely so, here houses may be regarded as clusters of units or elements, where the elements are persons living in the houses. In this case the units comprising the population of persons can be identified only with the help of houses. Similarly, the population of fields growing a certain crop in a certain tehsil/taluka can be identified with the help of villages in which they are located with the help of cultivators who cultivate them. In a livestock population, the elements may be animals and a household which has a number of animals can be considered as a cluster of animals e.g. in estimating milk production of bovines, it will be more convenient to select all the animals in a household as constituting a sampling unit. Thus the units of the population in which we are ultimately interested, are called the elements and the group of units used for identifying them are termed as the clusters. Usually we deal in practice such type of population where units can be identified by means of groups (called clusters) easily. To obtain a sample of elements from such a population a suitable number of clusters are first selected and then all the elements in the selected clusters constitute the sample. Such a procedure of getting the sample is known as cluster sampling.

1.2 The clusters can be selected with equal or unequal probability with or without replacement. Now if instead of including all the elements in the selected clusters in the sample, only a fraction of them is included in the sample and studied, the procedure is known as sub-sampling or two stage sampling. The procedure can be generalised to more than two stages and then it is called Multi-stage sampling.

1.3 Sometimes, clusters refer to the natural groups of units, e.g. villages, households etc. If natural clusters do not exist then they may be formed artificially with the help of a well defined criterion or procedure. Cluster sampling can also be used along with sub-sampling in a survey with or without stratification e.g. in two stage sampling, the primary sampling units may be clusters of two or more villages rather than individual villages. Similarly, the clusters of a certain number of households rather than individual households may be the second stage unit. In this situation, a village or a household itself becomes an element of the population of villages/households.

1.4 As mentioned earlier, the choice of sampling unit is very important in sample surveys. The choice of sampling unit will depend upon the type of population, variation within clusters and size of the cluster. It will also depend upon the cost involved in travelling between and within clusters.

1.5 Further, for a given sample size, clustering of units would decrease the efficiency of the sampling design in terms of variance of estimators compared to simple random sampling unless formation of clusters is done in such a way that the elements within a cluster differ very much in their character value but are similar between cluster. But if efficiency is considered in terms of total cost involved than cluster sampling would be possibly much cheaper than simple random sampling. The smaller the variation between the clusters, the greater will be efficiency of cluster sampling. The choice of an efficient sampling design should not only take into consideration the variance of the estimator but also the cost aspect. This is one of the principal reason why cluster sampling is used in sample surveys so frequently.

CHAPTER - II

REVIEW OF LITERATURE

2.1 Efficiency of cluster sampling: As has already been stated, the efficiency of cluster sampling will depend upon the choice of sampling unit, variation within cluster and size of the cluster.

Suppose a population of NM units is divided in to N mutually exclusive clusters of M units each and are one cluster is selected with equal probability for estimating the population mean

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i \quad , \quad (\bar{Y}_i = \frac{1}{M} \sum_{j=1}^{M} Y_{ij})..(2.1.1)$$

where Y_{ij} is the value of the j th unit in the i th cluster. An unbiased estimator of the population mean $(\bar{Y})$ is the sample cluster mean which is given by the formula

$$\hat{\bar{Y}}_c = \frac{1}{M} \sum_{j=1}^{M} Y_{ij} \qquad ....(2.1.2)$$

Where the subscript 'c' denotes that the estimator is based on a cluster sample. Its variance would be given by

$$V(\hat{\bar{Y}}_c) = \frac{1}{N} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2 = \sigma_b^2 \qquad (2.1.3)$$

where σ_b^2 stands for between cluster variance.

2.2 For studying the efficiency of cluster sampling as compared to simple random sampling, the variance of the sample mean $(\hat{\bar{Y}})$ based on M units drawn from NM units with equal probability with replacement would be

$$V(\hat{\bar{Y}}_r) = \frac{1}{M} \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y})^2 = \frac{\sigma^2}{M}$$

.......(2.2.1)

Where the subscript 'r' denotes simple random sampling and σ^2 is the total variance (when the deviation is taken from the population mean value ignoring the clusters for each unit in the population). The efficiency of cluster sampling as compared to simple random sampling with replacement would be

$$E_c = \frac{V(\hat{\bar{Y}}_r)}{V(\hat{\bar{Y}}_c)} = \frac{1}{M} \frac{\sigma^2}{\sigma_b^2} \quad ...(2.2.2)$$

From this it can be seen that cluster sampling will be mire efficient than simple random sampling with replacement only if the total variance (σ^2) is greater than M times the between cluster variance (σ_b^2) and this suggests that for the cluster sampling to be more efficient, the clusters should be formed such that variation between clusters means is very very small while that within clusters should be as large as possible.

2.3 Hansen and Hurwitz (1942) expressed the efficiency of cluster sampling in terms of intraclass correlation coefficient ρ between units within clusters.

$$\sigma_b^2 = \frac{\sigma^2}{M} \left[1+(M-1)\rho\right] \text{ where } \rho = \frac{2 \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{j'}^{M} (Y_{ij}-\bar{Y})(Y_{ij'}-\bar{Y})}{NM(M-1)\sigma^2} \qquad (2.3.1)$$

The efficiency of cluster sampling compared to SRS with replacement is then given by

$$E_c = \frac{1}{1+(M-1)\rho} \qquad (2.3.2.)$$

which shows that cluster sampling will be more efficient than simple random sampling only if ρ is negative. If M is equal to 1, then both are equally efficient. They also pointed out that usually ρ is postive and as M increases ρ decreases but the rate of decrease in ρ will be less as compared to rate of increase in M. Hence with increase in cluster size the variance of cluster sampling increases. And in case of sampling without replacement,

$$E_c = \frac{M(N-1)}{NM-1} \left[\frac{1}{1+(M-1)\rho}\right] \qquad (2.3.3.)$$

If N is large than both are equally efficient.

2.4 A number of attempts have been made to express the variance of the estimated characteristic from a sample of clusters of any size, given the variance of an equivalent sample of clusters of a given size as an explicit function of the cluster size. Smith (1938) proposed the relationship:

$$\sigma_b^2 = \frac{\sigma^2}{M^g} \qquad (2.4.4.)$$

Mahalanobis (1940) and Jessen (1942) suggested the relationship $\sigma_w^2 = aM^b$ $(b>0)$ where a and b are constant and $\sigma_w^2 = \sigma^2 - \sigma_b^2$, since these relationships are not exact and based on empirical investigation, so they should be verified before using for finding the optimum cluster size. Hendricks (1944) pointed out that the law suggested by Mahalanobis and Jessen may not hold good for large size clusters as correlation between the elements within a group approaches zero as the group become infinitely large. Asthana (1950) also reached the same conclusion, by fitting the law to describe the mean square within clusters, for a large number of villages when entire village was considered to be sampling unit.

2.5 However, the efficiency of any design has to be judged along with the consideration of cost incirred. The simplest and more appropriate cost function suggested by Jessen (1942) on empirical basis $C = C_1 nM + C_2\sqrt{n}$ (2.5.1.)

This is the cost equation for a sample of n clusters of M units, where C = total cost of the survey. C_1 = cost of enumeration per element including the travel cost from one element to another within cluster. C_2 = cost of travelling a unit distance between clusters.

2.6 From this sample, ignoring fpc, the variance of the mean per element will be

$$V(\hat{Y}) = \frac{1}{n}\left[\sigma^2 - (M-1)aM^{b-1}\right] \qquad (2.6.1)$$

For the optimum size of unit or optimum sampling unit M is to be determined, at a fixed level of cost C, such that $V(\hat{Y})$ is minimum. Jessen solved the problem experimentally by collecting data on farms using the sampling unit of different sizes for two given levels of cost. He worked out relative standard error for all sampling units and determined optimum sampling unit.

Cochran (1948,63) attempted the problem of choosing M and n algebraically so that the variance of the estimate is minimum for a fixed cost by minimising $\phi = V + \lambda(C_1 nM + C_2 n^{1/2} - C_0)$ where λ is the lagrangian multiplier which is undetermined and M, n denotes the cluster size and sample size respectively. Equations obtained on differentiation and equating to zero do not give explicit values of n and M and the method to solve them is trial and error.

2.7 Singh (1956) compared the efficiency of cluster sampling with that of sub-sampling procedure. He reached the conclusion that in many surveys, where the travelling cost between two second stage units is high single stage cluster (cluster consisting of second stage units) sampling may be preferred. He also observed that there was not much advantage in adopting two stage cluster sampling (selection of clusters of s.s.u's at second stage) over that of sub-sampling procedure. He also considered the case in which average travelling cost between two s.s.u.'s in the sample was relatively much larger than the average cost of selecting and locating the unit, field identification, listing etc. per second stage unit and where intraclass correlation was not high and suggested in such a case two stage cluster sampling may be more efficient than sub-sampling for a fixed cost of the survey.

2.8 In case of clusters of unequal sizes, the theory of cluster sampling becomes more complicated. In such cases, the methods of selection and estimation available provide either biased estimates with smaller variances or unbiased estimate with rather larger variances. In case of unequal clusters, let M_i denotes the size of the i^{th} cluster. For the sampling of clusters with equal probability and without replacement, the ratio estimator is

$$\hat{\bar{Y}}_2 = \sum_{i=1}^{n} M_i \bar{Y}_{Mi} \Big/ \sum_{i=1}^{n} M_i \qquad (2.8.1)$$

It is a biased estimate but may not be seriously biased for large n.
Similarly the estimator based on simple mean of the cluster means

$$\hat{\bar{Y}}_1 = \sum_{i=1}^{n} \bar{Y}_{Mi} / n \qquad (2.8.2)$$

This estimate of population mean may be seriously biased if cluster means and cluster sizes are correlated.

Further, if the sizes of all clusters, M_i's are known in advance then bias may be avoided by selecting a sample of clusters with probabilities proportional to cluster sizes and with replacement.

Horvitz & Thompson (1952), Yates & Grundy (1953) Hartley & Rao (1962) also suggested the method for selecting the clusters with p.p.s. and without replacement, where bias is negligible for large population. However, these methods are not useful for $n > 2$.

Sampford (1962), suggested the method of "Inverse Sampling" with p.p.s. (where sampling is done with p.p.s. with replacement). He also showed experimentally that this method gives unbiased estimate with lower variance comparative to other methods. However, in this method, knowledge of the sizes of all clusters is required which may not be available and also it is a costly method.

2.9 In sample surveys, where the list of elements is available but because using element as a sampling unit is either in convenekent or costly or both. In such situations clusters have to be formed artifically, though efficiency will become less but convenient and cheap.

Sethi (1965) discussed the problem of forming clusters of two units when information on the study variable (y) is available from a population for estimating the population mean or total. In this case units are arranged in increasing order with respect to value of the variable divided by the expected number of occurrences. He suggested that the clusters should be formed in such a way that given the method of estimation and the loss function, the risk is minimised. For a population consisting of 2N units, he suggested the clustering procedure as follows:

After arranging the units in increasing order, group the units which are equidistant from the ends. In this study cost aspect has not been considered while suggesting the criterion for forming clusters.

In such a case the cost of forming the clusters may be high and also the travelling cost between the elements of a cluster will be high unless these happen to be quite close to each other. Further for clusters of size greater than 2 the method becomes laborious.

2.9 Panse, Singh and Murty (1964,66); Singh, Murty and Goel (1970) and Singh, Rajgopalan and Maini (1970) have mentioned the use of method of formation of clusters which lead to the selection of overlapping clusters. These workers, however, for the purpose of estimation presumed that the clusters are non-overlapping. Murty (1967) also discussed overlapping clusters but did not discuss about the nature of the bias. Goel (1973) has also discussed the case of overlapping clusters and he has studied the nature of the bias being introduced with the use of overlapping clusters and efficiency of certain system of cluster sampling in case of overlapping clusters.

From the literature on cluster sampling, it is clear that several workers have discussed different methods of formation of clusters, but generally these methods lead to the formation of overlapping clusters. In this investigation, the problem of formation of non-overlapping clusters will be discussed and different estimators for estimating population parameter with their efficiencies will be studied.

CHAPTER - III

MATERIAL AND METHODS

3.1 Concept of Non-overlapping clusters (clusters without common elements)

Non-overlapping clusters impose a very serious restriction on cluster sampling in the sense that a unit can be included in only one cluster. On the other hand in the overlapping clusters an element may be included in more than one cluster. In this investigation, it is presumed that the list of units in the population is available and these units form distinct and non-overlapping groups (every element in the population belongs to one and only one group) and the criterion for forming clusters and selecting a sample of clusters is such that once an element is selected in a cluster then it cannot form part of any other cluster. Similarly, the second and subsequent clusters are also distinct and non-overlapping, sampling being done without replacement.

3.2 On the formation of clusters : Usually, we find situations in which cluster sampling is applied to the populations of natural clusters because list of ultimate sampling units is not available easily and the cost of preparing such a frame is very high. On the other hand, the list of clusters is generally available so clusters may be taken as the sampling units. However, sometimes, cluster sampling is also used in situations where the list of elements is available but because selection of elements as sampling units is inconvenient and costly. In such a case we may have to first form clusters of elements artifically from the list of elements available. In the formation of clusters, following points are important to consider.

(i) whether clusters are of equal or unequal sizes.

(ii) what should be the cluster size

(iii) how cluster formation should take place (what should be the criterion for forming clusters)

size and composition of the clusters are important because they are related to the cost of the survey and efficiency of the estimates.

When we conduct a survey, we generally try to avoid the use of unequal clusters because of many theoretical and practical complications. The sample size in terms of elements is a random variable which introduces further complications in the calculation of variance of the sample etimate.

The choice of size of cluster is a typical problem. A number of workers have studied this problem assuming that a unique criterion for forming clusters is available but in many cases such criterion may not be available. The optimum size of the cluster depends upon the composition of clusters and hence on the criterion for forming clusters.

3.3 It is therefore obvious that, when clusters do not exist in a natural way and cluster sampling procedure isto be adopted, then clusters have to be formed artifically, the choice of a suitable criterion for forming clusters becomes an important aspect. For this purpose the nature of the population size and shape of the elements, the type of association between elements of the population etc. will have to be considered. This is useful when the cost of travelling between the elements is an important factor.

3.4 Further, a criterion for formation of clusters may or may not lead to clusters which are non-overlapping. In case the clusters formed are non-overlapping, we draw a random sample of non-overlapping clusters and computes the inclusion probabilities of the various elements in the sample. An attempt will then be made to build up an unbiased estimate of the population total/mean. Alternatively, we may build up a simpler estimate to avoid laborious computations but it may not be unbiased and we would then determine the nature and extent of the bias to judge the efficiency of the estimate. These various aspects have been studied and discussed in the following sections.

3.5 Sampling Procedure: Many authors have suggested various methods of formation of overlapping clusters. Goel (1973) has also given a number of examples of formation and use of overlapping clusters. But the main advantage of the method of selecting non-overlapping clusters as described here is that we need not form all the clusters before selecting the sample. We form only as many clusters as are to be actually included in the sample.

Before discussing the sampling procedure, we make some important assumptions about the population. We assume that the population consists of distinct and non-overlapping groups. Also the number of elements belonging to a particular group are assumed known.

The procedure of selection of the elements and formation of clusters in the sample is as follows:
Let N denotes the total number of units in the population. One element is selected randomly from the population, since the group to which the selected unit belongs is known, another unit is selcted from the remaining units which belong to the same group to which the earlier unit belongs to form a cluster of two elements. Another cluster of two units is similarly formed by selecting one unit at random from the remaining (N-2) units in the population and then combining this unit with the another unit selected from the remaining units in the group to which the earlier unit belongs. This procedure can be continued till we get the desired number of clusters in the sample say 'n'. These 'n' clusters will be a random sample out of all the possible clusters say N', which can be formed if the procedure mentioned above is continued until all the population elements are exhausted. These N' clusters need not necessarily be equal. Clusters can be of unequal sizes depending upon the number of units in the various groups in the population. In this study a cluster consists of two elements and sometimes one element may also form a cluster in case it is selected and in the last remaining unit in the group.

In this way, N', non-overlapping clusters are obtained where N' is such that $\sum_{i=1}^{N'} M_i = N$ where M_i is the number of elements in the i th cluter (M_i takes the values two or one in the present study). The set of N' clusters so constructed is itself a random sample out of a very large number of similar sets which can be constructed by selection of different elements of the population at various stages of sampling. As already stated, the clusters in any one of these sets being non-overlapping will not have any element in common while the clusters of different sets will be overlapping.

This procedure of cluster sampling is equivalent to sampling with unequal probabilities and without replacement even though the first element of each selected cluster is selected with equal probability. In this sampling procedure the probability of selection of first unit is $\frac{1}{N}$, but as soon as it is selected another unit has to combine with it to form cluster of two units which has to be selected randomly from the group to which earlier selected unit belongs which means for the selection of second units in all other groups are given zero probability of selection. Using this criterion, the next unit will be selected from the rest of (N-2) units and thus a non-zero probability of selection is associated with every element for inclusion in the sample.

This technique of clustering after sampling may be extended to clusters of larger size (when cluster size is three or more), the cluster size will become more variable since it can assume all values less than the given size. The determination of the selection/inclusion probabilities of various units in the sample would obviously become more cumbersome and involved. However, with the availability and application of electronic computer, such complications may not be difficult to overcome.

In this study we shall discuss two cases (i) when the sample consists of two clusters (n = 2) and (ii) when the sample consists of three clusters (n = 3). The main objective of the study is to examine whether it would be possible to build up an unbiased estimate of the parameter under study using cluster sampling or if a simpler estimator is available which however, may not be unbiased so that we have to determine the nature and extent of bias in the estimate. The first step therefore is to compute the probabilities of selection/inclusion of various units in the sample. One method for this would be the procedure of enumeration in which all possible number of cases in which a sample of a given size can bo drawn are enumerated and the number of cases in which a particular unit is occurring which provide the frequency of occurrence of a unit in the sample. This method would, however, be very cumbersome and laborious when sample size is large. For small sample size, say 2, however, it would not be very difficult to adopt this approach.

Let us consider a group containing 'a' units. Every unit in this group may be associated with any one of the (a-1) remaining units. Let the total number of associating units be denoted by p(p=a-1), then p will take the values 1,2,3,...,k, where k is the number of the associating units in the largest group. Also more than one group may have the same 'p' value.

3.6 <u>Selection of Clusters: Case I:</u> When n = 2,

where 'n' denotes the number of clusters selected in the sample, when a sample of two clusters is taken from the population of N elements using the sampling procedure as described above, two cases arise (a) both clusters are from the same group, or (b) from different groups.

These two cases are mutually exclusive cases and account for the all possible ways in which a sample of two clusters can be taken.

(a) This event can occur only when $p \geq 2$. If $p = 1$, for the group to which the first selected unit belongs, we shall necessarily have case (b). When $p=2$, then a sample of two (unequal) clusters may become possible, e.g. in the group consists of elements (X, Y, Z), Then (X,Y) and (Z) may be reg arded as a sample of two unequal clusters. Similarly (X, Z)(Y), (Y,Z) (X) may also constitute the sample. However, these three samples are same as the units occuring in these are same.

As mentioned earlier, for determining the probability of selection/inclusion of a unit in the sample, we require the frequency of occurrence of a unit in different samples for which the method of enumeration of all possible samples containing different elements was used. Consider a population consists of k distinct types of groups and let n_p denotes the number of groups for which p (the number of associating units) value is same.

The number of cases in which a sample of two clusters can be drawn from a group is given as follows:

when

$p = 2 \nmid 1$

$$p = 3 : \quad \frac{(p-1)(p-2)}{2}$$

$$p = 4 : \quad \tfrac{1}{2}\left[(p-1)(p-2)+2(p-2)(p-3)\right]$$

$$p = 5 : \quad \tfrac{1}{2}\left[(p-1)(p-2)+2(p-2)(p-3)+3(p-3)(p-4)\right]$$

And in general for p = 3,4,5,...,k formula becomes

$$\tfrac{1}{2}\left[(p-1)(p-2)+2(p-2)(p-3)+3(p-3)(p-4)+\ldots\ldots(k-2)\{p-(k-2)\}\{p-(k-1)\}\right]$$

further which may be put in the form $\binom{p+1}{4}$

Hence, the total number of cases when the sample of two clusters belong to the same group will be given by

$$\tfrac{1}{2}\sum_{p=3}^{k}\left[(p-1)(p-2)+2(p-2)(p-3)+\ldots\ldots(k-2)\{p-(k-2)\}\{p-(k-1)\}\right]n_p+n_2$$

.......(3.6.1)

(b) The number of clusters which can be formed in a group with 'p' associating units is given by $\frac{p(p+1)}{2}$. The total number of clusters in the whole population would be given by

$$M = \sum_{p=1}^{k} \frac{p(p+1)}{2}\, n_p = \sum_{p=1}^{k} M_p\, n_p$$

where M_p denotes the number of clusters which can be formed within a group.

In case the two clusters in the sample come from different groups in the population, the number of samples of two clusters (any one of which may come) for a given group of p associating units is $\frac{p(p+1)}{2}\left[M - \frac{p(p+1)}{2}\right]$.

Hence the total number of possible samples in the population is

$$\frac{1}{2}\left[\sum_{p=1}^{k}\left\{\frac{p(p+1)}{2}\right\}\left\{M-\frac{p(p+1)}{2}\right\} n_p\right] \qquad (3.6.2)$$

Taking both the cases 'a' and 'b', the total number of possible samples are given by the formula:

$$\frac{1}{2}\left[\sum_{p=1}^{k}\left\{\frac{p(p+1)}{2}\right\}\left\{M-\frac{p(p+1)}{2}\right\} n_p\right] + \frac{1}{2}\sum_{p=3}^{k}\Big[(p-1)(p-2)+2(p-2)(p-3)$$

$$......+(k-2)\{p-(k-2)\}\{p-(k-1)\}\Big] n_p + n_2 \qquad .....(3.6.3)$$

For computation of inclusion probabilities (πp's) for various units in the population, it becomes necessary to study the relative frequencies of occurrence of different units in thepopulation. It is obvious that units belonging to the same type of groups (for which p values is same) will be havingsame values of (πp's).

In, case (a), the frequency of occurrence of a unit in a group is given by the formula

for $p = 3;$ $\frac{1}{2}\left[(p-1)(p-2)\right]$

$p = 4;$ $\frac{1}{2}\left[(p-1)(p-2)+(p-2)(p-3)\right]$

$p = 5;$ $\frac{1}{2}\left[(p-1)(p-2)+(p-2)(p-3)+(p-3)(p-4)\right]$

And in general for p = 3,4,5......k, formula becomes

$$\frac{1}{2}\left[(p-1)(p-2)+(p-2)(p-3)+......\{p-(k-2)\}\{p-(k-1)\}\right] \qquad ....(3.6.4)$$

For simplicity, which may be written as $\binom{p}{3}$

And also in case (b) the freq$\overset{ue}{\wedge}$ncy of occurrence of a unit which belongs to a group with its value p will be as follows:

$$\left[M - \frac{p(p+1)}{2} \right] p \qquad (3.6.5)$$

Thus combining the frequencies of different units in these two cases, the total number of possible samples containing a unit are obtained which provides the frequency of occurrence of a unit in a sample of two (non-overlapping) clusters. Formula is

$$f_p = \left[M - \frac{p(p+1)}{2} \right] p + \frac{1}{2}\left[(p-1)(p-2) + (p-2)(p-3) + \cdots\cdots \cdots\cdots \{p-(k-2)\}\{p-(k-1)\} \right] \qquad (3.6.6.)$$

when p=2, then formula becomes

$$f_2 = \left[\sum_{p=1}^{k} \frac{p(p+1)}{2} n_p - \frac{2(2+1)}{2} \right] 2 + 1$$

Hence we have got a unique formula for finding relative frequencies of different type of units whatever the value of 'p' is and hence (π_p's) can be computed.

The probability of inclusion (π_p's) of different units can be used to arrive at the average sample size (n') in terms of number of elements.

$$\sum_{p=1}^{k} \pi_p N_p = n' \text{ (average sample size)} \qquad (3.6.7)$$

where N_p: denotes the number of units in the population corresponding to p as the number of associating units.

3.7 <u>Case II:</u> when n = 3 (sample consists of three non-overlapping clusters)

In this case also, we proceed similarly as in case of n =2. There are three different ways of obtaining the sample of three clusters. Summing over all the three cases, we get the total number of ways in which a sample of three clusters can be drawn. The different cases are as follows:

(a) When all three clusters in the sample are drawn from different groups.

(b) Two clusters are selected from the same group and third one from the another group.

(c) All three clusters are selected from the same group.

The formulae and expression for sample of two clusters (n=2) have already been given. Proceeding similarly, formulae and expressions for sample size of three clusters (n=3) can be derived as follows:

(a) The sample of three clusters may be looked upon as one cluster drawn from any group in the population and the other two clusters from the remaining population. Thus the expression for the number of cases in which a sample of three clusters can be drawn from different groups would be as follows:

$$\frac{1}{3}\sum_{p=1}^{k}\left[\frac{p(p+1)}{2}\, n_p \left\{\begin{array}{l}\text{The number of cases in which a}\\ \text{sample of two clusters can be}\\ \text{drawn from different groups from}\\ \text{the remaining population (exclud-}\\ \text{ing) the group from which the fir}\\ \text{cluster has already been drawn).}\end{array}\right\}\right.$$

$$= \frac{1}{6} \sum_{p=1}^{k} \left[\frac{p(p+1)}{2} n_p \left\{ \sum_{p'=1}^{k} \frac{p'(p'+1)}{2} n_{p'} \left(M - \frac{p(p+1)}{2} - \frac{p'(p'+1)}{2} \right) \right\} \right] \quad (3.7.1$$

where 'p' denotes the number of associating units in the remaining population.

(b) The number of ways in which two clusters can be selected from the same group has already been given and is reproduced below:

$$s_p = \frac{1}{2} \sum_{p=3}^{k} \left[(p-1)(p-2)+2(p-2)(p-3)+.....+(k-2) \{p-(k-2)\}\{p-(k-1)\} \right] \quad (3.7.2$$

for $p = 3,4,5,....,k$

Thus, when two clusters are selected from the same group and one cluster from another group, the formula becomes

$$\sum_{p=1}^{k} s_p \left[M - \frac{p(p+1)}{2} \right] n_p \qquad (3.7.3)$$

where $s_p = 0$ for $p = 1$

$= 1;$ $p = 2$

$= 1;$ $p = 3$

There are also two other situations, when one cluster is selected from one group and other two clusters from the same group in the remaining population. However, these also lead to the same number of cases although the expressions are slightly different. The expression for the number of cases is given by

$$\sum_{p=1}^{k} \left[\frac{p(p+1)}{2} \left(\sum_{p=1}^{k} n_p s_p - s_p \right) n_p \right]$$

(c) In case all the three clusters are drawn from the same group, the expression may be derived as follows;

for $p = 6$; $\frac{1}{2}\left[(p-3)(p-4)+4(p-4)(p-5)\right]$

$p = 7$; $\frac{1}{2}\left[(p-3)(p-4)+4(p-4)(p-5)+10(p-5)(p-6)\right]$

$p = 8$; $\frac{1}{2}\left[(p-3)(p-4)+4(p-4)(p-5)+10(p-5)(p-6)+20(p-6)(p-7)\right]$

In general for $p = 6, 7, 8, \dots, k$, it becomes

$$\left[\binom{2}{2}\binom{p-3}{2} + \left\{\binom{2}{2} + \binom{3}{2}\right\}\binom{p-4}{2} + \left\{\binom{2}{2} + \binom{3}{2} + \binom{4}{2}\right\}\binom{p-5}{2} + \dots\dots\right.$$

$$\left. \dots + \binom{p-(k-3)}{2}\binom{p-(k-2)}{2} \right] n_p$$

Also it can be put in a simplified way $\binom{p+1}{6}$

and for $p = 4$; the number of cases become one

$p = 5$; again one

where n_p denotes the number of groups in the population with p value.

Hence combining all the three cases we get the total number of possible cases in which a sample of three (non-overlapping) cluster can be selected. Tho formula is given by

$$\frac{1}{6} \sum_{p=1}^{k} \left[\frac{p(p+1)}{2} n_p \left\{ \sum_{p'=1}^{k} \frac{p'(p'+1)}{2} n_{p'} \left(M - \frac{p(p+1)}{2} - \frac{p'(p'+1)}{2} \right) \right\} \right]$$

$$+ \sum_{p=1}^{k} s_p \left[M - \frac{p(p+1)}{2} \right] n_p + \left[\binom{2}{2}\binom{p-3}{2} + \left\{\binom{2}{2} + \binom{3}{2}\right\}\binom{p-4}{2} + \dots \right] n_p$$

$$\dots\dots\dots\dots\dots\dots(3.7.4)$$

In order to calculate the inclusion probabilities ($\wedge$p's) for different units in the population, we have to determine their frequencies of occurrence in the sample of three clusters. The frequency of any unit can be obtained by combining over the above three methods of obtaining the sample.

(a) In this case, let the groups be numbered as 1,2,3...,t i.e. 't' is the number of groups in the population and p takes the value 1,2,3,....,k. Also p_q' and p_1' denote the number of associating units in the q^{th} and 1^{th} group in the remaining population (exluding the group from which first cluster has been selected). Thus the frequency of an element belonging to the r^{th} group say with p as the number of associating units will be given by the formula

$$p_r\left[\sum_{q\neq r=1}^{t}\left(\sum_{1>q\neq r}^{t}\left\{\frac{p_q'(p_q'+1)}{2}\right\}\left\{\frac{p_1'(p_1'+1)}{2}\right\}\right)\right] \qquad (3.7.5)$$

Where $p_q' = 1,2,3....,k$

$p_1' = 1,2,3,.....,k$

(b) In this case, the frequency of an element which belongs to the group with value p (say) is

$$p\left[n_2' + \frac{1}{2}\sum_{p'=3}^{k}\left[(p'-1)(p'-2)+2(p'-2)(p'-3)+3(p'-3)(p'-4)+ ......\right.\right.$$

$$\left.\left.(k-2)\{p'-(k-2)\}\{p'-(k-1)\}\right]n_p'\right] \qquad (3.7.60$$

where n_p' and p' denotes the number of groups with number of associating units as p' in the remaining population (excluding the required groups for which frequency has to be determined).

(c). When all the three clusters selected in the sample come from the same group. This case arises only when $p \geqslant 4$,

when $p = 4$; the frequency of an element in this group is one (cluster of unequal sizes are formed).

$p = 5$; The frequency of an element in this group is also one,

$p = 6$; $\frac{1}{2}\left[(p-3)(p-4)+3(p-4)(p-5)\right]$

$p = 7$; $\frac{1}{2}\left[(p-3)(p-4)+3(p-4)(p-5)+6(p-5)(p-6)\right]$

$p = 8$; $\frac{1}{2}\left[(p-3)(p-4)+3(p-4)(p-5)+6(p-5)(p-6)+10(p-6)(p-7)\right]$

In general for $p = 6,7,8,\ldots,k$ formula becomes

$$\frac{1}{2}\left[(p-3)(p-4)+3(p-4)(p-5)+6(p-5)(p-6)+10(p-6)(p-7)+\ldots(k-4)^{\text{th}}\right]\text{term}$$

which is same as $\binom{p}{5}$

Thus, combining these above three cases, we get the frequency of a unit in the sample of three clusters as follows:

$$f_p = p\left[n_2'+\frac{1}{2}\sum_{p'=3}^{k}\left[(p'-1)(p'-2)+2(p'-2)(p'-3)+\ldots(k-2)\{p'-(k-2)\}\{p'-(k-1)\}\times n_p'\right]\right.$$

$$+\,p_r\left[\sum_{q\neq r=1}^{t}\left(\sum_{l>q\neq r}^{t}\left\{\frac{p_q'(p_q'+1)}{2}\right\}\left\{\frac{p_l'(p_l'+1)}{2}\right\}\right)\right]$$

$$+\frac{1}{2}\left[(p-3)(p-4)+3(p-4)(p-5)+6(p-5)(p-6)+10(p-6)(p-7)+\ldots\ldots\right] \quad (3.7.7$$

where p equals to $1,2,3,\ldots\ldots,k$ and other notations have their usual meanings.

Thus knowing relative frequencies of occurrence of a unit in a sample of three clusters and knowing the number of all possible cases in which a sample of three clusters can be drawn, the probabilities of inclusion (πp's) can be computed easily.

Having computed the probabilities of inclusion of different units in the sample. In the following discussions, the different estimates for the estimation of population parameter have been studied. Further, bias in estimation and variances of different estimates have been discussed also.

3.8 Estimation of population parameter (Total/Mean)

Let us consider a finite population consisting of N elements and let y be the character under study. Let y_i denote the value of the i th unit (i=1,2,3,..,N). These N elements may be grouped into a number of distinct and non-overlapping groups. It is desired to estimate the population total y Y which is given by

$$Y = \sum_{i=1}^{N} y_i$$

This can be estimated either from a sample of elements or from a sample of clusters of a given size.

Unbiased Estimate of population total Y

If inclusion probabilities π_i's (i=1,2,3....,N) are known, when a sample of n clusters each of size 'm' is taken, an unbiased estimate of the population total is given by

$$\hat{Y}_1 = \sum_{i=1}^{s} \frac{y_i}{\pi_i}$$

which is nothing but Horvitz Thompson estimate where 's' is the total number of elements selected in the sample. It may be noted that although cluster size m is 2, but it may sometimes takes the value 1 and as such the number of elements in the sample of size 's' will be either nm or nm-1.

It can be easily seen that $\hat{Y}_1 (= \sum_{i=1}^{s} \frac{y_i}{\pi_i})$ is an unbiased estimate of population total being a standard formula, and its variance is given by

$$V(\hat{Y}_1) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2 \sum_{i<}^{N} \sum_{j=1}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \quad \pi_j} y_i \, y_j \qquad (3.8.1$$

where π_{ij} denotes the probability of inclusion of i th and j th units simultaneously in a sample. The estimate of the variance is given by

$$v(\hat{Y}_1) = \sum_{i=1}^{s} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i<}^{s} \sum_{j=1}^{s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \; \pi_j \; \pi_{ij}} y_i y_j$$

........(3.8.2)

This is an unbiased estimate of the variance. There are, however, a few drawbacks of this estimate. It does not reduce to zero when all $z_i' = \frac{\text{[redacted]}_i}{m'} = \frac{y_i}{\pi_i}$ are equal in which case thevariance is necessarily zero. Also it may assume negative valtes for some samples.

3.9 Other estimates and their Variances:

Although the above procedure provides an unbiased estimate of the parameter (total/mean), it can be seen that computations of probabilities of inclusion (π_i's) of different units in the sample are quite laborious and cumbersome as already stated and hence computation of estimate and its variance become complicated.

Therefore instead of considering these π_i's as such in the estimate, one may consider some function of sample size say $\left(\frac{n'}{N}\right)$ where n' is average sample size which can be determined easily when all possible samples are known with their sizes. In our case of study, some samples may be of size nm and remaining of nm-1, where 'm' is supposed to take the value '2' so the total number of elements in all possible samples can be obtained by summing over these samples, and if this number is divided by the number of possible samples then n'(average sample size) is known. Thus n' may be calculated easily as compared to the calculation of π_i's

Thus, the estimate of say $(\hat{Y}_2)$ of population total becomes

$$\hat{Y}_2 = \frac{N}{n'} \sum_{i=1}^{s} y_i \qquad (3.9.1)$$

This estimate is however biased and its variance can be obtained as follows:

Let us define a random variable α_i (i=1,2,3,....,N) such that

$\alpha_i = 1$, if y_i is included in a sample of size 's' with probability π_i.

$= 0$, otherwise

$$E(\alpha_i) = 1\pi_i + 0(1-\pi_i) = \pi_i$$

and $V(\alpha_i) = E\left[\alpha_i - E(\alpha_i)\right]^2 = E(\alpha_i^2) - \left[E(\alpha_i)\right]^2 = \pi_i - \pi_i^2$

$$V(\alpha_i) = \pi_i\ (1-\pi_i)$$

and Cov. $(\alpha_i\,\alpha_j) = E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j) = (\pi_{ij} - \pi_i\,\pi_j)$

The expression for $(\hat{Y}_2)$ may be written as

$$\hat{Y}_2 = \frac{N}{n'} \sum_{i=1}^{N} \alpha_i \quad y_i$$

$$V(\hat{Y}_2) = V\left[\frac{N}{n'} \quad \sum_{i=1}^{N} \alpha_i \quad y_i\right]$$

$$= \frac{N^2}{n'^2}\left[\sum_{i=1}^{N} V(\alpha_i)y_i^2 + 2\sum_{i<j}^{N}\sum^{N} \text{Cov.}(\alpha_i\,\alpha_j)\, y_i \quad y_j\right]$$

$$V(\hat{Y}_2) = \frac{N^2}{n'^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)y_i^2 + 2\sum_{i<j}^{N}\sum^{N}(\pi_{ij}-\pi_i\pi_j)y_i\ y_j\right] \quad (3.9.2$$

and its estimate of variance is given by

$$v(\hat{Y}_2) = \frac{N^2}{n'^2}\left[\sum_{i=1}^{s} (1-\pi_i)\, y_i^2 + 2\sum_{i<j}^{s}\sum^{s} \frac{(\pi_{ij} - \pi_i\,\pi_j)}{\pi_{ij}}\; y_iy_j\right]$$

and π_{ij}'s are determined by calculating the relative frequencies of occurrence of (i,j)th unit (i th & j th unit simultaneously) in different samples and total no. of samples. ......(3.9.3)

This estimate of the variance is unbiased since $E\left[v(\hat{Y}_2)\right] = V(\hat{Y}_2)$ is true which may be shown as follows:

Let us define a variable α_i $(i=1,2,3,...,N)$ such that

$\alpha_i = 1$ if y_i is included in the sample with probability π_i

$= 0$ otherwise

so $E(\alpha_i) = \pi_i$, $v(\hat{Y}_2)$ may be written as follows

$$v(\hat{Y}_2) = \frac{N^2}{n'^2}\left[\sum_{i=1}^{N}(1-\pi_i)\alpha_i y_i^2 + 2\sum^{N}_{i<j}\sum^{N}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\alpha_i\alpha_j\, y_i y_j\right]$$

taking expectation both the sides, we get

$$E\left[v(\hat{Y}_2)\right] = E\left[\frac{N^2}{n'^2}\left\{\sum_{i=1}^{N}(1-\pi_i)\alpha_i\, y_i^2 + 2\sum^{N}_{i<j}\sum^{N}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\alpha_i\alpha_j y_i y_j\right\}\right]$$

Thus, $E\left[v(\hat{Y}_2)\right] = V(\hat{Y}_2)$

The above expressions for $V(\hat{Y}_2)$ and $v(\hat{Y}_2)$ derived above are general expressions. However, for evaluating the estimate of the variance, we require the value of each of π_i's for units selected in the sample, which as already stated, might be cumbersome and laborious process when the sample size or the population is large. A simple way out would be to ascribe equal probability to all the units in the sample i.e. $\pi_i = \frac{n}{N}$ which simplified substantially the evaluation of estimation of the variance. However, the estimate becomes biased since in this case $E\left[v(\hat{Y}_2)\right] \neq V(\hat{Y}_2)$.

The estimate of the variance under this assumption takes the form

$$v(\hat{Y}_2) = \frac{N}{n'^2}\left[(N-n')\sum_{i=1}^{s} y_i^2 - 2\,\frac{(N-n')}{(n'-1)}\sum_{i<j}^{s}\sum^{s} y_i\,y_j\right] \qquad (3.9.4$$

The bias in this estimate would be small if π_i's do not vary much and are close to $\frac{n'}{N}$.

As already stated, the estimate of thepopulation total is biased when π_i's are replaced by n'/N, this may be seen from the following.

$$\hat{Y}_2 = \frac{N}{n'}\sum_{i=1}^{s} y_i \text{ which may be written as}$$

$$\hat{Y}_2 = \frac{N}{n'}\sum_{i=1}^{N} \alpha_i\, y_i \text{ where } \alpha_i \ (i=1,2,\ldots\ldots,N) \text{ is defined earlier.}$$

taking expectation, we get

$$E(\hat{Y}_2) = E\left[\frac{N}{n'}\sum_{i=1}^{N}\alpha_i\, y_i\right] = \frac{N}{n'}\sum_{i=1}^{N} E(\alpha_i)\, y_i$$

$$\text{i.e. } E(Y_2) = \frac{N}{n'}\sum_{i=1}^{N} y_i\pi_i \neq Y$$

Thus, it would be of interest to study the bias in the estimate to determine its efficiency, the bias may be worked out as follows:

$$\text{Bias }(\hat{Y}_2) = E(\hat{Y}_2) - Y$$

$$= \frac{N}{n'}\sum_{i=1}^{N} y_i\pi_i - \sum_{i=1}^{N} y_i$$

$$= \sum_{i=1}^{N} y_i\left(\frac{N}{n'}\pi_i - 1\right)$$

$$\text{Bias } (\hat{Y}_2) = \frac{1}{n'} \sum_{i=1}^{N} y_i \, (N\pi_i - n') \qquad (3.9.5)$$

And estimate of bias may be as follows

Since; Bias $(\hat{Y}_2) = E(\hat{Y}_2) - Y$

i.e. Bias $(\hat{Y}_2) = E(\hat{Y}_2) - E(\hat{Y}_1) = E(\hat{Y}_2 - \hat{Y}_1)$ where $\hat{Y}_1 = \sum_{i=1}^{s} \frac{y_i}{\pi_i}$

Hence, est. Bias $(\hat{Y}_2) = \hat{Y}_2 - \hat{Y}_1$

i.e. est. Bias $(\hat{Y}_2) = \frac{N}{n'} \sum_{i=1}^{s} y_i - \sum_{i=1}^{s} \frac{y_i}{\pi_i}$

$$\text{est. Bias } (\hat{Y}_2) = \frac{1}{n'} \sum_{i=1}^{s} y_i \left(N - \frac{n'}{\pi_i} \right) \qquad (3.9.6)$$

The magnitude of the bias depends upon the nature of values of π_i's. From the above expression, it is clear that if π_i's are approximately equal to n'/N, then it approaches to zero almost. However, if π_i's are proportional to y_i's (observation value) then bias will be significant, which may or may not be true in our case of study.

As may be seen, the calculation of n', which provides the average sample size, require the knowledge of all possible samples with their sizes. A further simplification may therefore be made, by replacing n' (average sample size) by s, the size of the sample drawn already. Another estimate of the population total may then be build up as follows:

$$\hat{Y}_3 = \frac{N}{s} \sum_{i=1}^{s} y_i$$

This is also a biased estimate of the population total, since

$E\left[(\hat{Y}_3)\right] \neq Y$,as may be seen from the following.

The sample size 's' becomes a random variable and takes the value nm or nm-1. Let denote the probability of the two values by $p(n*m)$ and $p(nm-1)$ respectively, and let π_i' and π_i'' denote the inclusion probabilities in the two groups respectively for every element ($i=1,2,3...,N$).

Then $E(\hat{Y}_3) = E\left[\frac{N}{s} \sum_{i=1}^{s} y_i\right]$

which may be written as

$$E(\hat{Y}_3)=E_1\left[\frac{N}{nm} \sum_{i=1}^{nm} y_i\right]p(nm)+E_2\left[\frac{N}{nm-1} \sum_{i=1}^{nm-1} y_i\right]p(nm-1)$$

Here also, similarly α_i and β_i are defined as earlier and their expected values are computed, on simplification

$$E(\hat{Y}_3) = \frac{N}{nm}\, p(nm) \sum_{i=1}^{N} y_i \pi_i' + \frac{N}{nm-1}\, p(nm-1) \sum_{i=1}^{N} y_i \pi_i''$$

which shows $E(\hat{Y}_3) \neq Y$.

As may be seen, when all possible samples are of the same size, (nm or nm-1) then $(\hat{Y}_3)$ is the same as $(\hat{Y}_2)$, since the average sample size is also the same, in fact, this will provide the estimate of population total as in case of SRS and its estimate of variance is given by

$$v(\hat{Y}_3) = \frac{N}{s^2}\left[(N-s) \sum_{i=1}^{s} y_i^2 + 2 \frac{(N-1)}{(s-1)} \sum_{i<j}^{s} \sum^{s} y_i\, y_j \right] \quad (3.9.7)$$

The estimate of bias in case of $\hat{Y}_3 \left(= \frac{N}{s} \sum_{i=1}^{s} y_i\right)$ is

given by:

$$\text{Est Bias } (\hat{Y}_3) = \hat{Y}_3 - \hat{Y}_1 = \frac{N}{s} \sum_{i=1}^{s} y_i - \sum_{i=1}^{s} \frac{y_i}{\pi_i}$$

$$\text{Hence, est. Bias } (\hat{Y}_3) = \sum_{i=1}^{s} y_i \left(\frac{N}{s} - \frac{1}{\pi_i} \right) \qquad (3.9.8)$$

3.10 Comparison with sampling with equal probabilities and without replacement

Let us have a population of N distinct elements, to estimate population total (Y), a sample of s units is selected out of N units with equal probability and without replacement. The estimat of Y is given by,

$$\hat{Y}_{srs} = \frac{N}{s} \sum_{i=1}^{s} y_i$$

and $V(\hat{Y}_{srs}) = \frac{N(N-s)}{s} \sigma^2$ where $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2$

and $v(\hat{Y}_{srs}) = \frac{N(N-s)}{s} \hat{\sigma}^2$

Comparison of efficiency of different estimates

If T_1 and T_2 are two unbiased estimates of a parameter θ, then efficiency of T_2 in relation to that of T_1 is defined as

$$E = \frac{V(T_1)}{V(T_2)}$$

And if T_1 and T_2 are biased estimates then their relative efficienc is given by

$$E = \frac{\text{Mean Square Errar}(T_1)}{\text{Mean Square Error}(T_2)} = \frac{V(T_1) + \text{Bias}^2 (T_1)}{V(T_2) + \text{Bias}^2 (T_2)}$$

In our case of study, $(\hat{Y}_1)$ is an unbiased estimate and $\hat{Y}_{srs}$ also provides an unbiased estimate ofpopulation total.

Thus, Relative efficiency of $(\hat{Y}_1)$ compared to srs. $= \dfrac{V(\hat{Y}_{srs})}{V(\hat{Y}_1)}$

$$\text{i.e.} \quad = \frac{N(N-s)\,\sigma^2}{s} \Big/ \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i}\, y_i^2 + 2 \sum_{i<j}^{N}\sum^{N} \frac{(\pi_{ij}-\pi_i\pi_j)\,y_i y_j}{\pi_i\ \pi_j}$$

And it is well known that estimate $(\hat{Y}_1)$ will be more efficient estimate of Y compared to srs, if inclusion probabilities (π_i) are proportional to the some measure of the size of the units (property of Horwitz estimate) but in our case this condition does not hold strictly. but as already, stated, π_i's may not be easy to compute in case of large population or large sample size etc.

We consider the biased estimate $\hat{Y}_2 (= \frac{N}{n'} \sum_{i=1}^{s} y_i)$ and compare its efficiency with srs$(\hat{Y}_{srs})$ which is unbiased estimate, it being assumed that the estimates are based on the same sample of size 's'.

Relative efficiency of $(\hat{Y}_2)$ compared to srs $= \dfrac{V(\hat{Y}_{srs})}{V(\hat{Y}_2) + \text{Bias}^2\ (\hat{Y}_2)}$

$$= \frac{N(N-s)}{s}\,\sigma^2 \Big/ \frac{N^2}{n'^2}\left[\sum_{i=1}^{N} \pi_i\,(1-\pi_i)\,y_i^2 + 2\sum_{i<j}^{N}\sum^{N}(\pi_{ij}-\pi_i\pi_j)\,y_i y_j\right] + \left[\frac{1}{n'} \sum_{i=1}^{N} y_i\,(N\pi_i - n')\right]^2$$

The estimates of variances of $\hat{y}_2$ and $\hat{Y}_{srs}$ were calculated for a no. of cases with diffent types of data in which $\hat{y}_2$ turns out to be more efficient than $\hat{Y}_{srs}$, hence it may be shown from the above expression of relative efficiency, the estimate $\hat{Y}_2$ is more efficient than $(\hat{Y}_{srs})$ when π_i's for different units in the sample are replaced by the function of sample size n'/N.

Also efficiency of two sampling procedures may be compared if T_1 and T_2 are two estimates based on two different samples obtained according to two different sampling procedures. These are illustrated with the help of survey data in following discussions.

CHAPTER-IV

RESULTS AND DISCUSSION

4.1 In situations where cluster sampling procedures are adopted and the clusters do not exist in a natural way, a system has been suggested in Chapter III for forming non-overlapping clusters, the estimates of the population parameter, the bias in the estimates and their efficiencies for sampling with unequal probabilities (althougth they are not unequal prob. in usual sense, the prob. will vary only very slightly) and the sampling with equal probabilities were also discussed. These will be illustrated with the help of data of survey conducted in Punjab during 1971. The sampled cultivators in a district were treated as the population. The number of cultivators selected in a village under the survey were found to vary from village to village. For the purpose of this study, therefore, the villages with varying number of cultivators were taken as the group of variable sizes in the population. The procedure developed earlier are applied to this populationfor estimating the average doses of application of N, P and K. In our case, the population of 24 cultivators with their respective quantities of N, P and K applied to a field are given in Appendix I.

Thus the population consists of 24 cultivators coming from five villages (which may be considered as 5 groups of units) and a sample of 4 cultivators is drawn using samplig procedure described earlier. The sampe values are given below.

Figures are in kg/Acre.

	p = 2		p = 6	
N	46.0	46.0	40.0	46.0
P	9.0	6.6666	8.4	8.0
K	11.25	12.0	12.0	11.6666

Also, s = 4 and $n' = 3.998$

The estimates of N, P and K, bias and the estimate of their variances for different estimation procedures are presented in Table I.

From the results in the Table I, it is clear that when probability of inclusion (π_i's) are taken into consideration for estimation of population mean (Using Horvitz Thompson estimate, $\hat{\bar{Y}}_1$, which is an unbiased estimate) then this estimate turns out to be an over-estimate of the population averages of N, P and K. The estimates of variances are also fairly high. The two estimates of $\hat{\bar{Y}}_2$ (when π_i's are replaced by n'/N) and $\hat{\bar{Y}}_3$ (when π_i's are replaced by $\frac{s}{N}$) differ slightly in their estimated values of population mean. Further, in our case, the estimate $\hat{\bar{Y}}_3$ (when π_i are replaced by s/N) becomes srs estimate. The estimates of variances in case of $\hat{\bar{Y}}_2$ is less than that given by srs and Horvitz estimate for the same sample size, which has been found true in the no. of cases and one of the case is discussed here, which implies the estimate $\hat{\bar{Y}}_2$ is more efficient than srs and Horvitz estimate. The magnitude of bias as estimated from the sample in this case is about 20 per cent.

Thus, when non-overlapping clusters are formed using the criterion mentioned earlier and a sample of say 'n' clusters (which may be unequal) is drawn, then the population parameter can be estimated unbiasedly using Horvitz Thompson estimate and estimate of variance may be computed provided the probabilities of inclusion of various units in the sample are known. However, this estimate is efficient only if π_i's are proportional to the values of elements selected in the sample. If π_i's are replaced by some function of sample size say $\frac{n'}{N}$, which simplifies the estimation procedure, the estimate of variance is lower and therefore this estimate is more efficient than the Horvitz Thompson estimate. Further, this estimate (using $\frac{n'}{N}$) is also more efficient than srs.

Besides being most efficient, the cost aspects is also quite favourable which in any case is the main reason for resorting to cluster sampling. Further, in case two different samples are taken with the different sampling procedures, then this sampling scheme proves to be more efficient.

APPENDIX I

State: Punjab Season: Kharif Year: 1971

Population of 24 Cultivators in 5 villages with fertilizer N,P & K

I			II			III			IV			V		
N	P	K	N	P	K	N	P	K	N	P	K	N	P	K
46.0	8.0	1.875	40.0	9.0	11.25	48.5	8.0	11.25	46.0	8.0	11.25	46.0	8.0	10.0
46.0	8.0	12.0	46.0	9.0	11.25	46.0	8.0	11.6666	40.0	7.5	11.25	46.0	8.0	11.6666
46.0	8.0	11.875	46.0	6.666	12.0	46.0	8.0	12.5	40.0	8.0	12. 5	46.0	8.0	14.0
46.0	7.5	12.5				40.0	8.4	12.0				20.0	8.0	12.5
46.0	8.0	12.5				46.0	8.0	11.6666				46.0	10.0	12.5
						40.0	8.5	11.25				46.0	8.0	11.25
						46.0	8.0	12. 0						

TABLE I

Comparative study of different estimates, bias and estimates of their variances

Name of Fertilizer	Population values Y	$\hat{\bar{Y}}_1$	$\hat{\bar{Y}}_2$	$\hat{\bar{Y}}_3 = \hat{\bar{Y}}_{srs}$	Est. Bias$(\hat{\bar{Y}}_2)$	$v(\hat{\bar{Y}}_1)$	$v(\hat{\bar{Y}}_2)$	$v(\hat{\bar{Y}}_3) = v(\hat{\bar{Y}}_{srs})$	%S.E. $(\hat{\bar{Y}}_1)$	%S.E. $(\hat{\bar{Y}}_2)$	%S.E.$(\hat{\bar{Y}}_3)$ = %S.E.$(\hat{\bar{Y}}_{srs})$
N	43.770	57.07	44.5222	44.5	-12.5478	495.55	1.6019	1.875	39.0	2.84	3.027
P	8.107	10.08	8.0207	8.0166	-2.0593	12.56	0.195	0.204	35.16	5.508	5.634
K	11.440	14.82	11.7350	11.729	-3.085	28.89	0.0073	0.0264	36.27	0.7286	1.385

N.B. Where $\hat{\bar{Y}}_1$ is $= \frac{1}{N} \sum_{i=1}^{s} \frac{y_i}{\pi_i}$

$$\hat{\bar{Y}}_2 = \frac{N}{n'} \cdot \frac{1}{N} \sum_{i=1}^{s} y_i$$

$$\hat{\bar{Y}}_3 = \frac{N}{s} \cdot \frac{1}{N} \sum_{i=1}^{s} y_i$$

SUMMARY

Cluster sampling is frequently used in sample surveys on account of cost, convenience and efficiency consideration Usually we come across situations in which cluster sampling is applied to the populations of natural clusters because list of elements is not available easily. However, sometimes natural clusters do not exist and the list of elements is available but because using element as a sampling unit is either inconvenient or costly. In such situations, clusters have to be formed artifichlly considering the important points like cluster size, criterion for forming clusters etc. Further, a criterion may lead to clusters which are overlapping or non-overlapping. Several workers have suggested the method of forming clusters but generally they lead to over-lapping clusters.

In this investigation, a criterion for forming non-overlapping clusters (clusters may be unequal) has been suggested where we need not form all the clusters before selecting the sample, we form only as many clusters as are to be actually included in the sample. The method of enumeration is used for computations of probabilities of inclusion (π_i's) of different units selected in the sample in which all possible samples are enumerated and the relative frequencies of occurrence of different units in the sample. Further, it has been shown in this dissertation, knowing π_i's, the population parameter can be estimated unbiasedly using Horvitz Thompson estimate. In this case, the estimate of variance is high. Further,

if λ_i's are replaced by some function of sample size say $\frac{n'}{N}$, which simplifies the estimation procedure and the estimate of variance is also lower in this case compared to srs and Horvitz estimate, although the estimate obtained in this way is a biased estimate. The magnitude of bias is also not significant, which proves the estimate obtained in this way is more efficient and the sampling scheme is fruitful.

REFERENCES

1. Cochran, W.G. (1963) "Sampling Techniques", Second edition, John Wiley & Sons, Inc. New York, London

2. Goel, B.B.P.S. (1973) "Efficiency of certain systems of cluster sampling and its application". Unpublished Thesis for Ph.D, I.C.A.R., New Delhi.

3. Hartley, H.O. and Rao, J.N.K. (1962) "Sampling with unequal probabilities and without replacement". Ann. Math. Stat. Vol.33 pp. 350-374.

4. Sampford, M.R. (1962) "Methods of cluster sampling with and without replacement for clusters of unequal sizes", Biometrika, Vol.49 pp.27-40.

5. Singh, D. (1956) "On efficiency of cluster sampling". Jour. Ind. Soc. of Agri. Stat. Vol.8 pp.45

6. Sukhatme, P.V. and Sukhatme, B.V. (1970) "Sampling Theory of surveys with applications". Asia Publishing House, Bombay.

7. Murty. M.N. (1967) "Sampling theory and Methods" Statistic Publishing House, Calcutta.

8. Sethi. V.K. (1965) "On optimum pairing of units" Sankhya (B) Vol.27 PP. 315-320

9. Murty, V.V.R. and Goel B.B.R.S. (1970) "Monograph on estimation of milk production" I.C.A.R. New Delhi.

10. Hansen M.H and Hurwitz W.N. (1942) "Relative efficiencies of various sampling units in population enquiries" Jour. Amer. Stat. Assoc. vol 37 PP. 89-94

11. Mahalanobis P.C. (1940) "A sample survey for acreage under Jute in Bengal" Sankhya Vol.4 PP. 511-530