SOME PROBLEMS OF ESTIMATION IN SAMPLING FROM FINITE POPULATIONS

By

M. N. MURTHY

July 1961
Indian Statistical Institute
Calcutta - 35.

P R E F A C E

In this thesis the results of the research work done by the author in the field of sample surveys are presented. Some problems of estimation in sampling from finite populations were taken up for research. A brief summary of this thesis is given below.

In chapter 1, the author has developed a generalized theory of getting unbiased estimators for a certain class of parameters in sampling from finite populations. The class of parameters considered are those which can be expressed as the sum of single-valued set functions defined over a class of sets of units belonging to a finite population. A technique of generating unbiased estimators for this class of parameters is given for any sample design. This is of importance, since so far unbiased estimators for any sample design have been suggested on 'a priori' and intuitive considerations and not as a result of generating technique. It is of interest to note that a particular general estimator given by the generating technique happens to include, as particular cases, most of the estimators commonly used in practice and this estimator may be taken as a 'reasonable estimator' whenever there is doubt as to which estimator is to be used.

The technique of interpenetrating sub-samples, introduced by Prof. P.C.Mahalanobis as long back as 1938, has been shown to have tremendous possibilities in estimating the bias of a certain class of non-linear parametric functions. In chapter 2 is given a technique of getting (almost) unbiased estimators based on independent interpenetrating sub-sample estimates for parametric functions which can

be expressed as non-linear functions of parameters which can be unbiasedly estimated using the technique given in chapter 1.

In chapter 3, it is shown that in case of sampling from a finite population without replacement, corresponding to any estimator based on the order of selection of the units in the sample (ordered extimator) there exists a more efficient estimator which ignores the order of selection of the units in the sample (unordered estimator). The ordered estimators suggested by different authors have been improved upon using this technique.

The technique of getting (almost)unbiased estimators developed in chapter 2 is applied to the case of ratio method of estimation in chapter 4. It is also shown that for any design the conventional biased ratio estimator and the unbiased ratio estimator based on a slightly modified design are equally efficient in large samples to second degree of approximation and that the latter is more efficient than the former to the fourth degree of approximation under certain assumptions.

A new method of estimation, termed 'product method of estimation' is introduced in chapter 5. This is similar to the ratio method of estimation and in fact this may be considered to supplement the ratio method. It is shown that in large samples, product method of estimation would be more efficient whenever ratio method of estimation is inefficient. The question of estimating the bias of estimators of

product of several parameters is also considered.

The efficiencies of short-cut methods of estimating variance and confidence interval are considered in chapter 6. It is shown that the loss of efficiency in estimating the variance and confidence interval on the basis of independent interpenetrating sub-sample estimates decreases more rapidly for initial increases in the number of sub-samples than for further increases. The estimator of variance built up from strata sub-sample estimates is formally shown to be more efficient than that based on the sub-sample estimates pooled over strata without any assumptions.

A procedure of determining the sample size, which can be considered as more rational than the conventional procedure, has been suggested and a specimen table showing the sample sizes for different situations is given.

The question of making a sample design self-weighting at field and tabulation stages has been investigated. The advantages and disadvantages of different methods of making sample design self-weighting are considered in chapter 7.

In chapter 8, an attempt is made to give a comprehensive treatment of the theory of survey errors including both sampling and non-sampling errors. The use of 'conditional approach' has considerably simplified the derivation of a number of results in this field. The technique of interpenetrating sub-samples (introduced by Prof. P. C. Mahalanobis) and its various uses in large scale surveys are discussed.

I wish to express my gratitude to Dr. C. R. Rao and Prof. D. B. Lahiri for their constant encouragement in the preparation of this thesis. I should like to thank Mr. K. Bhattacharyya for his help in typing the manuscript.

Calcutta,
July 1961.

M.N.Murthy

C O N T E N T S

Chapter 1

GENERALIZED UNBIASED ESTIMATION

1. INTRODUCTION

From times immemorial the concept of generalizing from a 'part' to the 'whole' has been used more or less subjectively in daily life. But not until the later half of the nineteenth century, objective methods of generalizing from a 'part' to the 'whole' seem to have received much attention. In this case two questions arise - (i) how to select the 'part' from the 'whole' and (ii) how to generalize from the selected part to the whole. The problem is one of finding that combination of selection and estimation procedures which would minimize the risk involved in generalizing from the part to the whole per unit of cost. Alternatively the problem may be viewed as one of finding that combination of selection and estimation procedures which would minimize the cost, ensuring at the same time a specified precision for the inference from the part to the whole.

The earlier developments in this field relate to the second question posed above and the result has been a fairly well developed theory of estimation and statistical inference based on the simplest of selection procedures, namely, equal probability sampling with replacement. Since the last two decades, a number of selection and

estimation procedures have been given in sampling from finite populations. Some attempts have been made to give generalized estimation procedures which would cover, as particular cases, the estimators commonly used in practice (Midzuno, 1950, Godambe, 1955, Murthy, Nanjamma and Sethi, 1959). Goodman (1953), Murthy (1957) and Basu (1958) have given techniques of improving upon certain types of estimators.

In this chapter, it is proposed to develop a generalized estimation procedure on the lines considered by Murthy, Nanjamma and Sethi (1959) and give a technique of generating estimators for any sample design. This technique is useful as it puts at our disposal a number of estimators. So far, the estimators in case of different sample designs have been suggested by various authors on 'a priori' and intuitive considerations and not as a result of a generating technique. Of course the problem of choice among different estimators remains and in most of the cases extensive empirical studies would be necessary to arrive at the best or near optimum estimators. This is because there does not exist a uniformly unbiased minimum variance estimator in the non-parametric sense in sampling from finite populations, as has been shown by Godambe (1955).

It is interesting to note that a particular general estimator given by the generating technique happens to include, as particular cases, most of the estimators commonly used in practice and this estimator may be taken as a 'reasonable' estimator whenever there is doubt

as to which estimator is to be used.

Before actually giving the generalized estimator and the technique of generating unbiased estimators for any sample design, it is proposed to define the terms 'sample design' and 'sampling scheme and to give the techniques of improving estimators mentioned above.

2. SAMPLE DESIGN AND SAMPLING SCHEME

A 'sample design' may be considered as the specification of all possible samples with their probabilities of selection or of the probabilities of inclusion of the units or combination of units in the sample. A 'sampling scheme' may be taken as the procedure of selection of the sample in stages by units or combination of units.

A sample design is completely specified by a list of possible samples including all possible permutations and repetitions of the units in the sample with their respective probabilities of selection. In such a case it can be seen that there is a unique sampling scheme which gives rise to the sample design, as has been formally shown by Hanumantha Rao (1960). Examples of partial specification of sample design are (i) specification of all possible samples without considering the permutations and repetitions of units in them with their respective probabilities and (ii) specification of the probabilities of inclusion of the units in the sample. In case of partially specified sample design there is no unique sampling scheme giving rise to the sample design. Instead there are a number of sampling schemes which satisfy the partial specifications of the sample design. In such cases the problem is to

find that selection procedure which minimises the variance of a given estimator.

For instance, in case of a partially specified sample design where the probabilities of inclusion of the units are specified, a simple selection procedure which is likely to be more efficient than the others is to select the sample with probability proportional to the probabilities of inclusion systematically after some suitable arrangement of the units (Murthy, 1960). The procedure consists in first obtaining the cumulated totals of π_i's, the probabilities of inclusion, ($C_i = C_{i-1} + \pi_i$, $i = 1, 2, \ldots, N$) and then selecting the units systematically with a random start from 0 to 1 and with 1 as the sampling interval using the cumulated totals of the π_i's.

3. IMPROVING OF ESTIMATORS

Three techniques of improving upon certain types of estimators have been suggested. Goodman (1953) has shown that if t is an unbiased estimator of θ and $V(t) = K\theta^2$, where K is known, then a more efficient estimator from the point of the risk function

$$\lambda(\theta)(t - \theta)^2, \quad \lambda(\theta) > 0 \tag{3.1}$$

is given by

$$t' = t/(K + 1). \tag{3.2}$$

An example of such a situation is provided by the estimator

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.3}$$

of σ^2 in case of a sample from a normal population. Since

$$V(s^2) = \frac{2}{n-1}\sigma^4$$

a better estimator of σ^2 is given by

$$s'^2 = \frac{1}{n+1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (3.4)$$

In case of sampling without replacement from finite populations, some estimators have been given by Das (1951) and Des Raj (1956), which depend on the order in which the units are selected in the sample. Murthy (1957) has shown that corresponding to any estimator based on the order of selection of the units, there exists a more efficient estimator which does not take into account the order of selection of the units in the sample. The former may be termed 'ordered estimator' and the latter 'unordered estimator'. The technique involved in improving the ordered estimator consists in taking the conditional expected value of the ordered estimator over all possible orders for a given unordered sample of units as the unordered estimator. For instance, in case of selecting 2 units with varying probabilities without replacement, one of the ordered estimators given by Des Raj is

$$\hat{Y}_o = \frac{1}{2}\left[\frac{y_i}{p_i}(1 + p_i) + \frac{y_j}{p_j}(1 - p_i)\right] \qquad (3.5)$$

where the order of selection of the units in the sample is (ij). The corresponding unordered estimator which is more efficient than the

above estimator is

$$\hat{Y}_u = \frac{1}{2 - p_i - p_j} \left[\frac{y_i}{p_i}(1 - p_j) + \frac{y_j}{p_j}(1 - p_i) \right] \quad (3.6)$$

This technique is considered in greater detail in chapter 3.

In case of sampling with replacement, Basu (1958) has shown that corresponding to any estimator which takes account of the number of repetitions of the units in the sample, there exists a more efficient estimator which is based only on the distinct units in the sample without taking into consideration the number of repetitions. The procedure of improving the estimator is similar to that explained earlier and consists in taking the conditional expected value of the estimator over all possible repetitions of the given set of distinct units. For instance in the case of simple random sampling with replacement, the sample mean based only on the distinct units in the sample is more efficient than the mean of the units in the sample including their repetitions. Similarly in case of sampling with probabilities proportional to a given measure of size with replacement, if $n - 1$ units turn out to be distinct, then the improved estimator is given by

$$\hat{Y}_d = \frac{1}{n} \left[(\bar{y}'/\bar{p}') + \sum_{i=1}^{d} (y'_i / p'_i) \right] \quad (3.7)$$

where $\bar{y}'$ and $\bar{p}'$ are the sample means of y and p based only on the distinct units.

The improved estimators obtained by using the above techniques are in general difficult to calculate except for certain particular cases

and their variance estimators are rather complicated.

4. PARAMETRIC FUNCTION

Let X denote a finite population of N units $(U_i, i = 1,2,\ldots, N)$ and let the vector $\underline{X}_i$ giving the values of a number of characteristics, be associated with the ith unit $(i = 1, 2, \ldots, N)$. Let A be the class of all sets 'a' whose elements belong to X. In such a set the same unit may or may not occur more than once. The class of all point sets and the class of all pairs of units belonging to X are examples of the class 'A'.

We shall confine ourselves to only such parametric functions (say F) which can be expressed as a sum of single valued set functions defined over the class 'A'. That is, if F is such a function, then

$$F = \sum_{a \in A} f(a) \qquad (4.1)$$

where $f(a)$ is a single valued set function defined over the class 'A' and $\sum_{a \in A}$ stands for the summation over all sets 'a' belonging to to the class 'A'.

For instance, the population total Y for a characteristic y can be expressed as F in (4.1) with 'a' as a point set (U_i) and $f(a)$ as Y_i. The population variance σ^2 can be expressed as F in (4.1) with 'a' as a set of two units $\{U_i, U_j\}$* and

$$f(a) = \frac{1}{N^2}(Y_i - Y_j)^2$$

* The curled brackets $\{\ \}$ are used to denote unordered sets, that is, $\{U_i, U_j\}$ and $\{U_j, U_i\}$ are the same.

Similarly the covariance between two characteristics x and y may be expressed as in (4.1) with 'a' as a pair of units $\{U_i, U_j\}$ and

$$f(a) = \frac{1}{N^2}(Y_i - Y_j)(X_i - X_j).$$

It may be noted that there may be a number of ways of expressing a parameter in the form (4.1). For instance the population total may as well be expressed as F with 'a' as a set of n units and

$$f(a) = \frac{n\bar{y}}{\binom{N-1}{n-1}}$$

where $\bar{y}$ is the mean of the values of y for the units in the set 'a'. In such cases, it is desirable to define the set 'a' over as small number of units of the population as possible and in what follows this is assumed to be so, unless otherwise stated.

5. SAMPLE SPACE AND ESTIMABILITY

Let w_{or} be a set of elements of X taking into account the permutations and repetitions of the units. The subscript 'o' and 'r' are used to denote the order and repetitions of the units respectively. Let W_{or} be the class of all such sets w_{or}. W_{or} may be considered the sample space. This together with a specified probability measure $P(w_{or})$ defined over it gives rise to a completely specified sample design. Before giving the generalized estimator for F, two of the three procedures of improving upon certain types of estimators mentioned in section 3 are considered in this context.

If $\hat{Y}_Y$ is an unbiased estimator which takes into account the

repetitions of the units in the sample, then a more efficient estimator which ignores the repetitions of the units and is based only on the distinct units in the sample w_{or} is given by

$$\hat{Y} = \sum_{w_r \to w} \hat{Y}_r P(w_r) / P(w) \qquad (5.1)$$

where w_r denotes a sample where the repetitions of the units are taken into account, but not the arrangement of the units in the sample, $P(w_r)$ and $P(w)$ are the probabilities of getting the samples w_r and w respectively, where w is the sample ignoring repetitions and arrangement of the units, and $\sum_{w_r \to w}$ denotes summation over all samples w_r corresponding to the sample w.

Similarly if $\hat{Y}_o$ is an unbiased estimator of Y based on the ordered sample w_o, then a more efficient estimator which is based on the unordered sample w corresponding to w_o is given by

$$\hat{Y} = \sum_{w_o \to w} \hat{Y}_o P(w_o) / P(w) \qquad (5.2)$$

$\sum_{w_o \to w}$ denotes summation over all possible ordered samples w_o corresponding to the unordered sample w.

For the sake of convenience, let us consider the case where the sample size is fixed. Let 's' be a set of units of X with or without repetitions and with or without arrangement of the units, as the case may be. The class of all such sets is the sample space S and let the probability of getting the sample be $P(s) > 0$. It may be noted

that 's' is being used here to denote w_{or}, w_o, w_r or w as the case may be. Similarly S stands for W_{or}, W_o, W_r or W as the case may be. Without loss of generality, it may be assumed that each sample 's' consists of at least one set 'a'. If this were not so, we can redefine the sample space and the probability measure such that this is so.

THEOREM 1 : The parametric function given in (4.1) can be estimated unbiasedly from the sample if and only if each set 'a' is contained in at least in one sample 's'.

Proof : That the condition is sufficient can easily be seen by considering the estimator

$$\hat{F} = \sum_{a \subset s} f(a)\, \phi\,(s, a) \,/\, P(s) \qquad (5.3)$$

where $\phi(s,a)$ is some function of 's' and 'a'. The necessary and sufficient condition for this estimator to be unbiased for F is that

$$\sum_{s \supset a} \phi\,(s,a) = 1 \qquad (5.4)$$

for $E(\hat{F}) = \sum_{s \in S} \sum_{a \subset s} f(a)\, \phi(s,a)$

$$= \sum_{a \in A} f(a) \sum_{s \supset a} \phi\,(s,a).$$

There are a number of functions satisfying (5.4), as can easily be verified. The necessity for the condition can be proved as follows. Suppose some set 'a' is not included in any sample. Then the coefficient of the corresponding $f(a)$ in $E(\hat{F})$ given above will be 0 and not 1. Hence the necessity for the condition.

THEOREM 2 : An unbiased estimator of the variance of the estimator $\hat{F}$ given in (5.3) can be obtained if and only if every set $(a \cup a')$ is contained in at least one sample 's'.

Proof : For estimation of the variance of $\hat{F}$ unbiasedly, it is sufficient to estimate F^2 unbiasedly, for

$$E(\hat{F}^2) = V(\hat{F}) + F^2$$

$$\text{i.e. Est. } V(\hat{F}) = \hat{F}^2 - \text{Est. } F^2.$$

F^2 can be expressed as

$$F^2 = \sum_{a \in A} \sum_{a' \in A} f(a)\, f(a').$$

From this it follows that an unbiased estimator of F^2 is possible if and only if every set $(a \cup a')$ is included in at least one sample 's' (cf. Theorem 1).

An unbiased estimator of $V(\hat{F})$ is given by

$$\hat{V}(\hat{F}) = \hat{F}^2 - \Big[\sum_{a \subset A} \sum_{a' \subset A} f(a)\, f(a')\, \psi(s, a \cup a') \Big] / P(s) \qquad (5.5)$$

where

$$\sum_{s \supset a \cup a'} \psi(s, a \cup a') = 1. \qquad (5.6)$$

If some supplementary information (say $g(a)$) related to $f(a)$ is available for all the sets in 'A' and is made use of at the selection stage, then it is desirable that whenever $g(a)$ is substituted for $f(a)$ in the estimator (5.3), we get the population parameter

$$G(a) = \sum_{a \in A} g(a).$$

In other words, it is desirable to find $\phi(s,a)$ in the estimator (5.3) such that

$$\sum_{s \supset a} \phi(s,a) = 1$$

and

$$\sum_{a \subset s} g(a)\, \phi(s,a) / P(s) = G \qquad (5.7)$$

for each sample 's'.

6. GENERALIZED UNBIASED ESTIMATOR

It is proposed to consider the estimator given in (5.1), namely,

$$\hat{F} = \sum_{a \subset s} f(a)\, \phi(s,a) / P(s) \qquad (6.1)$$

and to give a technique for generating unbiased estimators for any sample design. As has been shown earlier, $\hat{F}$ in (6.1) is unbiased for F if and only if

$$\sum_{s \supset a} \phi(s,a) = 1 \qquad (6.2)$$

One solution for (6.2) is given by

$$\phi(s,a) = 1/N(s \supset a) \qquad (6.3)$$

where $N(s \supset a)$ stands for the number of samples containing the set 'a'.

In that case the estimator is given by

$$\hat{F}_1 = \sum_{a \subset s} f(a) / N(s \supset a)\, P(s) \qquad (6.4)$$

where $\sum_{a \subset s}$ stands for summation over all sets 'a' included in the sample 's'.

THEOREM 3 : An unbiased estimator of P given in (4.1) is given by

$$\hat{P} = \sum_{a \subset s} f(a)\, P(s/E_a) \,/\, P(s) \qquad (6.5)$$

where E_a is any event concerning the occurrence of the set 'a' in the sample satisfying the following conditions:-

(i) the occurrence of E_a implies the selection of one $s \supset a$, and

(ii) the selection of 's' implies E_a for at least one $a \subset s$,

and $P(s/E_a)$ is the conditional probability of selecting the sample 's' given that the event E_a has occurred.

Proof : The theorem will be proved if we can show that the condition (6.2) is satisfied when

$$\phi(s,a) = P(s/E_a).$$

In this case the condition is satisfied, for

$$P(s/E_a) = P(s,E_a) \,/ \sum_{A \supset a} P(s, E_a),$$

where $P(s, E_a)$ is the probability of getting the sample 's' where the event E_a has occurred, and hence

$$\sum_{s \supset a} P(s/E_a) = 1.$$

Since $P(s/E_a)$ satisfies (6.2) for all specifications of the event E_a satisfying the conditions mentioned in Theorem 3, we can get many estimators by defining the event E_a in different ways. The event E_a may be termed the 'generating event' since it generates estimators. For instance, the following specifications of the event may be considered:-

(i) the set 'a' is included in the sample 's' - $E_{a\subset s}$,

(ii) the set 'a' occurs first in ordered sample - E_{a1},

(iii) the set 'a' occurs in the second position in the ordered sample - E_{a2},

and so on. The estimator corresponding to the events $E_{a\subset s}$, E_{a1} and E_{a2} are given by

$$\hat{F}_2 = \sum_{a\subset s} f(a) / \pi(a) \qquad (6.6)$$

where $\pi(a)$ is the probability of including the set 'a' in the sample, since

$$P(s/E_{a\subset s}) = P(s) / P(a\subset s),$$

$$\hat{F}_3 = \sum_{a\subset s} f(a)\, P(s/E_{a1}) / P(s) \qquad (6.7)$$

$$\hat{F}_4 = \sum_{a\subset s} f(a)\, P(s/E_{a2}) / P(s) \qquad (6.8)$$

The estimator $\hat{F}_i$ satisfies the desirable condition mentioned in (5.7) if $\pi(a)$ is made proportional to the supplementary information $y(a)$ related to $f(a)$. For substituting $g(a)$ for $f(a)$ in (6.6) in the case we get

$$\sum_{a\subset s} \frac{g(a)}{\pi(a)} = Q$$

since $\sum_{a\in A} \pi(a)$ is the number of sets 'a' occurring in sample 's'.

Similarly the estimator $\hat{F}_3$ satisfies this condition if one set 'a' is selected with probability proportional to $g(a)$ in the first draw. For in that case we get

$$\sum_{a\subset s} \frac{g(a)\, P(s/Ea)}{P(s)} = Q$$

by substituting $g(a)$ for $f(a)$ in (6.7).

It may be noted that for the estimator to be useful in practice, the generating event E_a should be so specified that it would be possible to calculate the conditional probability $P(s/E_a)$ from the information available about the population, the sample and the sampling scheme. For instance if the event is defined as the occurrence of the set 'a' such that $f(a) < f(a')$ for all $a' \subset s$, it is not possible to calculate the conditional probability $P(s/E_a)$ since the values of $f(a)$ are available only for the a's included in the sample 's'.

The technique of generating estimators given above shows that given any sampling scheme, one can derive a number of unbiased estimators by defining the generating event E_a in different ways and choose one of these taking into consideration cost and efficiency. In other words, this technique gives rise to a variety of estimators to choose from. As would become clear from the examples considered in later sections, the estimator obtained by considering the event E_a as the occurrence of the set 'a' first in the ordered sample includes as particular cases most of the estimators may be taken as a 'reasonable estimator' and used whenever there is doubt as to which estimator to use in particular situations. It may be noted that though the specification of the generating event takes into account the order of selection, the resulting estimator need not be an ordered estimator.

If the distinct units in the sample and the probabilities of selecting all possible unordered samples are given, then it will be possible to generate a number of estimators by

(i) specifying different sampling schemes giving rise to the partially specified sample design, and

(ii) defining the generating event in different ways.

To fix the ideas, let us consider a sample design specifying the unordered samples and their respective probabilities,

$$\{U_i, U_j\} \text{ and } \pi_{ij}$$

where $$\pi_{ij} = p_i p_j \left[\frac{1}{1 - p_i} + \frac{1}{1 - p_j} \right]$$

p_i's being less than 1 and that $\sum_{i=1}^{N} p_i = 1$.

Since the sampling scheme is unique for a completely specified sample design we get different sampling schemes by specifying different ways of distributing the probability of the unordered sample $\{U_i, U_j\}$ over the ordered samples (U_i, U_j) and (U_j, U_i). For instance, let

$$P_1(U_i, U_j) = p_i p_j / (1 - p_i) \text{ and } P_2(U_j, U_i) = p_i p_j / (1-p_j) \quad (6.9)$$

$$P_2(U_i, U_j) = \pi_{ij}/2 \text{ and } P_2(U_j, U_i) = \pi_{ij}/2 \quad (6.10)$$

In both the cases

$$P(U_i, U_j) + P(U_j, U_i) = \pi_{ij}$$

Let the generating event be the occurrence of the unit first in the ordered sample. For (6.9) $P(s/U_i 1) = p_j/(1-p_i)$ and the estimator is

$$\hat{Y}' = \frac{1}{2-p_i-p_j} \left[\frac{y_i}{p_i} (1-p_j) + \frac{y_j}{p_j} (1-p_i) \right] \quad (6.11)$$

whereas for the same generating event in case of (6.10), we get

$$\hat{Y}'' = \frac{y_i}{\pi_i} + \frac{y_j}{\pi_j}, \quad (\pi_i = \sum_{j \neq i} \pi_{ij}) \qquad (6.12)$$

Similarly other estimators may be obtained for different specifications of the generating event.

Another method of obtaining unbiased estimators using the above technique is to express the parametric function in the form (4.1) in a number of ways by adopting different definitions of the class 'A' of sets 'a'. For instance the population total Y can be expressed as

$$Y = \sum (m \bar{y}_m) / \binom{N-1}{m-1} \quad m = 1, 2, \ldots, N. \qquad (6.13)$$

where $\bar{y}_m$ is the mean based on a combination of m units from N units and $\sum$ stands for summation over all combinations of m units from N units.

Thus we see that if a sample of size n is taken, then n possibly different estimators can be obtained by taking different values of m (= 1, 2, ..., n) in (6.13). The estimators will be of the form

$$\hat{Y}_m = \sum_m (m \bar{y}_m) \phi(s,a) / \binom{N-1}{m-1} P(s), \quad (m = 1, 2, \ldots, n) \qquad (6.14)$$

where $\sum_m$ stands for summation over all combinations of m units from n units in the sample and 'a' stands for a set of m units.

To summarise, it may be noted that a number of unbiased estimators may be obtained from (6.5) by different specifications of

(i) sampling scheme,

(ii) generating event and

(iii) the set 'a' .

A number of estimators generated by the techniques explained in this section in case of different particular situations are considered in the later sections.

7. VARIANCE ESTIMATOR

In section 5 it has been shown that an unbiased estimator of the variance of $\hat{F}$ is given by

$$\hat{V}(\hat{F}) = \hat{F}^2 - \left[\sum_{a \subset s} \sum_{a' \subset s} f(a)\, f(a')\, \psi(s, a \cup a') \right] / P(s) \qquad (7.1)$$

where

$$\sum_{s \supset a \cup a'} \psi(s, a \cup a') = 1 \,. \qquad (7.2)$$

Now we give a technique of generating unbiased variance estimators by giving a procedure for obtaining a number of solutions of (7.2) To start with we may take

$$\psi(s, a \cup a') = 1/N(s \supset a \cup a') \qquad (7.3)$$

where $N(s \supset a \cup a')$ stands for the number of samples containing the set $(a \cup a')$ and this gives rise to the following unbiased estimator of the variance of $\hat{F}$,

$$\hat{V}_1(\hat{F}) = \hat{F}^2 - \left[\sum_{a \subset s} \sum_{a' \subset s} f(a) f(a') / N(s \supset a \cup a') P(s) \right] \qquad (7.4)$$

A number of estimators can be obtained by taking

$$\psi(s, a \cup a') = P(s/E_{a \cup a'}) \qquad (7.5)$$

where $P(s/E_{a \cup a'})$ is the conditional probability of getting the sample 's' given that the generating event concerning the occurrence of the set $a \cup a'$ in the sample has occurred. It can be easily verified that this satisfies (7.2), for

$$P(s/E_{a \cup a'}) = P(s, E_{a \cup a'}) \Big/ \sum_{s \supset a \cup a'} P(s, E_{a \cup a'}).$$

Taking up the illustration considered in section 6, we see that unbiased estimators of Y^2 for (6.9) and (6.10) are respectively given by

$$\frac{1}{2-p_i-p_j}\left[\frac{y_i^2}{p_i}(1-p_j)+\frac{y_j^2}{p_j}(1-p_i)\right]+2\frac{y_iy_j}{\pi_{ij}} \qquad (7.6)$$

and

$$\frac{y_i^2}{\pi_i}+\frac{y_j^2}{\pi_j}+2\frac{y_iy_j}{\pi_{ij}}. \qquad (7.7)$$

It may be noted that for each estimator, there are a number of unbiased variance estimators generated by the above technique.

As mentioned earlier it is proposed to consider in the next few sections the different estimators and their variance estimators this technique gives rise to when it is applied to some commonly used sampling schemes. For the sake of simplicity, only the question of estimating the population total Y of the characteristic y and in some cases estimation of the population variance σ^2 on the basis of a sample of fixed size with or without repetitions and with or

without arrangement of the units in the sample as the case may be is considered. Of course, this technique of generating estimators may be applied to get unbiased estimators of a number of other parametric functions, such as moments, which can be expressed in the form of F in (4.1).

8. SIMPLE RANDOM SAMPLING

8.1 With replacement scheme. Suppose a sample of n units is selected from a finite population of N units with equal probability with replacement and let y_i $(i = 1, 2, \ldots, n)$ be the value of the ith unit in the sample. Here two cases arise, (i) estimators which take into account the repetitions of the units in the sample and (ii) estimators based only the distinct units in the sample. As mentioned earlier, corresponding to any estimator of type (i), there exists a more efficient estimator of type (ii).

If 's' is a sample of n units selected with equal probability with replacement where the repetitions and arrangement of the units in the sample are taken into consideration, we have

$$P(s) = 1/N^n$$

$$N(s \ni i) = n N^{n-1}$$

with the interpretation that $N(s \ni i)$ is the number of repetitions of the ith unit in all possible samples,

$$P(s/E_{11}) = P(s/E_{12}) = \ldots = 1/n N^{n-1}.$$

Hence we find that the different specifications of the generating event considered above lead to the following estimator

$$\hat{Y} = N\bar{y} = N\left(\sum_{i=1}^{n} y_i\right)/n \qquad (8.1)$$

In this case $E_{i \subset s}$ is difficult to define, since repetitions of the units in the sample are taken into account in the estimator. Since

$$P(s/E_{ij1}) = 1/\binom{n}{2}N^{n-2},$$

an unbiased estimator of the variance of $\bar{y}$ is given by

$$\hat{V}(\bar{y}) = \hat{\bar{y}}^2 - \left[N^2 \sum_{i=1} \sum_{j>i} y_i y_j / \binom{n}{2}\right]$$

$$= N^2 s^2 / n \qquad (8.2)$$

where $s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)$.

Suppose we consider only estimators based on the distinct units in the sample. In this case, we have

$$P(s) = P(d)\, P(s/d) = P(d) / \binom{N}{d}$$

where $P(d)$ is the probability of getting d distinct units in the sample,

$$N(s \supset i) = N^n \pi(i)$$

where $\pi(i)$ is the probability of inclusion of the ith unit and is given by

$$\pi(i) = 1 - (1 - \tfrac{1}{N})^n$$

and $P(s/E_{i1}) = P(s/E_{i2}) = \ldots = P(d)\ P(s/d, E_{id}) = P(d) \,/\, \binom{N-1}{d-1}$.

The estimators are

$$\hat{Y}_1 = \hat{Y}_2 = d\,\bar{y}' \,/\, \left[\, 1 - (1 - \tfrac{1}{N})^n \,\right] \qquad (8.3)$$

and

$$\hat{Y}_3 = \hat{Y}_4 = \ldots = N\,\bar{y}' \qquad (8.4)$$

where $\bar{y}'$ is the mean based on the d distinct units. Incidentally the estimator (8.4) can be obtained from (8.1) by using the technique of improving estimators considered in section 3. The estimator (8.3) is the one suggested by Bodambe (1955). Since

$$P(s/E_{ij1}) = P(d)\ P(s/d, E_{ij1}) = P(d) \,/\, \binom{N-1}{n-1} \text{ for } i = j$$

$$= P(d) \,/\, \binom{N-2}{n-2} \text{ for } i \neq j$$

an unbiased estimator of variance of the estimator (8.4) is given by

$$V(N\,\bar{y}') = N^2 y'^2 - \left(\frac{N}{d} \sum_{i=1}^{d} y_i'^2 + 2\,\frac{N(N-1)}{n(n-1)} \sum_{i=1}^{d} \sum_{j>i} y_i' y_j' \right)$$

$$= N(N-d)\, s'^2/d \qquad (8.5)$$

where $s'^2 = \sum_{i=1}^{d} (y_i' - \bar{y}')^2 \,/\, (d-1)$.

8.2 Without replacement scheme. In sampling n units from a population of N units with equal probability without replacement, we get for an unordered sample 's'

$$P(s) = 1 / \binom{N}{n}$$

$$N(s \supset i) = \binom{N-1}{n-1}$$

$$\pi(i) = n/N$$

$$P(s/E_{im}) = 1/\binom{N-1}{n-1}, \quad (m = 1, 2, \ldots, n).$$

Here we see that all the estimators generated by different solutions of (6.2) considered above lead to the following estimator

$$N\bar{y} \tag{8.6}$$

where $\bar{y}$ is the sample mean. It can be easily verified that an unbiased estimator of the variance of this estimator is

$$\hat{V}(N\bar{y}) = N(N-n)s^2/n \tag{8.7}$$

where $s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$.

Since $\sigma^2 = \sum_{i=1}^{N} \sum_{j>i} (Y_i - Y_j)^2 / N^2$ and $P(s) = 1/\binom{N}{n}$,

$$P(s/E_{ij1}) = 1/\binom{N-2}{n-2}$$

an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \sum_{j>i} (y_i - y_j)^2 \, P(s/E_{ij1})/N^2 P(s)$$

$$= (N-1)\, s^2/N \tag{8.8}$$

where s^2 is as defined in (8.7).

9. VARYING PROBABILITY SAMPLING

9.1 With replacement scheme. Suppose n units are selected with probability proportional to a given measure of size with replacement where the probability of sampling the ith unit at each draw is P_i $(i = 1, 2, \ldots, N)$. Let s be an unordered sample of d distinct units obtained by the above procedure taking into account the repetitions of the units in the sample and let p_i be the probability of selecting the ith distinct unit in the sample 's'. In this case we get

$$P(s) = \frac{n!}{r_1!\, r_2! \ldots r_d!}\, p_1^{r_1}\, p_2^{r_2} \ldots p_d^{r_d}$$

where r_i is the number of repetitions of the ith unit in the sample,

$$P(s/E_{im}) = \frac{(n-1)!}{r_1!\, r_2!\, (r_i - 1)! \ldots r_d!}\, p_1^{r_1}\, p_2^{r_2}\, p_i^{r_i-1} \ldots p_d^{r_d}$$

$(m = 1,2,\ldots, n)$.

Hence we get the estimator

$$\hat{Y}_3 = \hat{Y}_4 = \ldots = \hat{Y}_n = \left(\frac{1}{n}\sum_{i=1}^{d} r_i y_i / p_i\right) \qquad (9.1)$$

An unbiased estimator of the variance of this estimator is given by

$$\hat{V}(\hat{Y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}\right)^2 \qquad (9.2)$$

since an unbiased estimator of Y^2 is in this case

$$\sum_{i=1}^{n}\sum_{j>i}\left(y_i y_j / p_i p_j\right) / \binom{n}{2}.$$

If we take the sample 's' as consisting of only the distinct units without taking into account repetitions or arrangement of the units in the sample, we get the estimators

$$\hat{Y}_2 = \sum_{i=1}^{d} y_i / [1 - (1-p_i)^n] \qquad (9.3)$$

since $\pi(i) = [1 - (1 - p_i)^n]$

and

$$\hat{Y}_3 = \sum_{i=1}^{d} y_i \, P(s/E_{i1})/P(s) \qquad (9.4)$$

The expression for the conditional probability $P(s/E_{i1})$ of getting the sample 's' given that the ith unit has been selected first is in general rather complicated. To illustrate the situation, let us consider the case where there are n-1 distinct units in the sample of n units selected with probability proportional to size with replacement. Then

$$P(s) = \frac{n!}{1!\,1! \ldots 1!\,2!} p_1 p_2 \cdots p_{n-1} \sum_{i=1}^{n-1} p_i$$

$$P(s/E_{im}) = \frac{(n-1)!}{1!\,1! \ldots 1!\,2!} p_1 p_2 \cdots p_{n-1} \frac{\left(\sum_{j \neq i}^{n-1} p_j \right)}{p_i}$$

$$+ \frac{(n-1)!}{1!\,1! \ldots 1!} p_1 p_2 \cdots p_{n-1}$$

$$= \frac{P(s)}{n} \left[\frac{1}{p_i} + 1 \Big/ \left(\sum_{i=1}^{n-1} p_i \right) \right]$$

Hence we get the estimator

$$\hat{Y}_m = \frac{1}{n} \left[\sum_{i=1}^{n-1} \frac{y_i}{p_i} + \frac{\sum_{i=1}^{n-1} y_i}{\sum_{i=1}^{n-1} p_i} \right] \qquad (9.5)$$

Incidentally this is the estimator obtained by improving the estimator (9.1) by using the technique given in section 3 when there are (n-1) distinct units in the sample.

9.2 Without replacement scheme. In sampling with varying probabilities without replacement, the generalised estimator gives rise to a number of different estimators for different specifications of the generating event E concerning the occurrence of the ith unit in the sample. For the sake of simplicity, let us consider the case where the sample size is 2.

Case (i) pps and pps of the remaining. Let the selected unordered sample be $\{U_i, U_j\}$ and let the initial probabilities for these units be p_i and p_j. In this case

$$P(s) = \pi_{ij} = p_i p_j (2 - p_i - p_j)/(1-p_i)(1-p_j)$$

$$N(s \ni i) = N-1$$

$$\pi(i) = \pi_i = \sum_{j \neq i} \pi_{ij}$$

$$P(s/i1) = p_j/(1-p_i)$$

$$P(s/i2) = p_i p_j / (1 - p_j)(\pi_i - p_i).$$

Unbiased estimators of Y corresponding to the different solutions of (6.2) considered above are given by

$$\hat{Y}_1 = \frac{y_i + y_j}{(N-1)\,\pi_{ij}} \qquad (9.6)$$

$$\hat{Y}_2 = \frac{y_i}{\pi_i} + \frac{y_j}{\pi_j} \qquad (9.7)$$

$$\hat{Y}_3 = \frac{1}{(2-p_i-p_j)}\left[\frac{y_i}{p_i}(1-p_j)+\frac{y_j}{p_j}(1-p_i)\right] \qquad (9.8)$$

$$\hat{Y}_4 = \frac{1}{(2-p_i-p_j)}\left[\frac{y_i}{(\pi_i-p_i)}(1-p_i)+\frac{y_j}{(\pi_j-p_j)}(1-p_j)\right] \qquad \ldots \quad (9.9)$$

It is of interest to note that (9.6) is a particular case of the estimator suggested by Midzuno (1950), (9.7) is the estimator suggested by Horvitz and Thompson (1952) and (9.8) was obtained by Murthy (1957) by unordering one of the ordered estimators considered by Des Raj (1956). The estimator (9.9) is a new estimator. Thus we see that all the estimators commonly used in varying probability sampling without replacement can be generated by the technique introduced here. There are two other estimators, namely,

$$\hat{Y}_5 = \tfrac{1}{2}(Y_1 + Y_3) \qquad (9.10)$$

$$\hat{Y}_6 = \tfrac{1}{2}(Y_3 + Y_4) \qquad (9.11)$$

which are obtained by unordering the ordered estimators

$$\hat{Y}' = \left\{\frac{y_i}{p_i} + \frac{1}{N-1}\,\frac{(1-p_i)}{p_i}\,\frac{y_j}{p_j}\right\} \qquad (9.12)$$

$$\hat{Y}'' = \left\{\frac{y_i}{P_1(i)} + \frac{y_j}{P_2(j)}\right\} \qquad (9.13)$$

where the ordered sample is (ij) and $P_1(i)$ and $P_2(j)$ are respectively the unconditional probabilities of getting ith unit in the first draw and the jth unit in the second draw. The estimator in (9.12)

was suggested by Das (1951) and that in (9.13) by Des Raj (1956). Here we see that the unordered estimators (9.10) and (9.11) corresponding to (9.12) and (9.13) are nothing but linear combinations of the estimators generated by the technique under consideration. One may as well consider other combinations of the estimators given in (9.6) to (9.9). That is, we may consider the estimators

$$\hat{Y}_7 = (Y_1 + Y_2) / 2 \qquad (9.14)$$

$$\hat{Y}_8 = (Y_1 + Y_4) / 2 \qquad (9.15)$$

$$\hat{Y}_9 = (Y_2 + Y_3) / 2 \qquad (9.16)$$

$$\hat{Y}_{10} = (Y_2 + Y_4) / 2 \qquad (9.17)$$

The estimators corresponding to (9.10) and (9.11) in the general case where the sample size is greater than 2, are derived in chapter 3.

It is of interest to note that of all the estimators given above only $\hat{Y}_5$ satisfies the desirable condition given in (5.7). For in that case, if we substitute p_i for y_i in the estimator, we get 1 which is the total of the initial probabilities of all the units in the population.

Unbiased estimators of the variances of the estimators given above can be obtained, if unbiased estimators of Y^2 are given. In this case estimators of

$$\sum_{i=1}^{N} Y_i^2$$

can be got by substituting y_i^2 and y_j^2 for y_i and y_j in the above expressions. An unbiased estimator of

$$2 \sum_{i=1}^{N} \sum_{j>i} Y_i Y_j$$

is given by

$$2y_i y_j / \pi_{ij} .$$

Thus we see that for each estimator given above there are a number of variance estimators. For instance,

$$\hat{V}_1(\hat{Y}_1) = \hat{Y}_1^2 - \left[\frac{y_i^2 + y_j^2}{(N-1)\pi_{ij}} + 2 \frac{y_i y_j}{\pi_{ij}} \right] \qquad (9.18)$$

$$\hat{V}_2(\hat{Y}_1) = \hat{Y}_1^2 - \left[\frac{y_i^2}{\pi_i} + \frac{y_j}{\pi_j} + 2 \frac{y_i y_j}{\pi_{ij}} \right], \text{ etc.}$$

Similarly we get

$$\hat{V}_1(\hat{Y}_2) = \hat{Y}_2^2 - \left[\frac{y_i^2 + y_j^2}{(N-1)\pi_{ij}} + 2 \frac{y_i y_j}{\pi_{ij}} \right]$$

$$\hat{V}_2(\hat{Y}_2) = \hat{Y}_2^2 - \left[\frac{y_i^2}{\pi_i} + \frac{y_j^2}{\pi_j} + 2 \frac{y_i y_j}{\pi_{ij}} \right]$$

$$= \frac{y_i^2}{\pi_i^2}(1 - \pi_i) + \frac{y_j^2}{\pi_j^2}(1 - \pi_j) + 2 \frac{y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j) \qquad (9.19)$$

which is the variance estimator proposed by Horvitz and Thompson (1952) for their estimator (9.7),

$$\hat{V}_3(\hat{Y}_3) = \hat{Y}_3^2 - \frac{1}{2-p_i-p_j} \left[\frac{y_i^2}{p_i}(1-p_i) + \frac{y_j^2}{p_j}(1-p_j) + \frac{y_i y_j}{\pi_{ij}} \right]$$

$$= \frac{(1-p_i)(1-p_j)(1-p_i-p_j)}{(2-p_i-p_j)^2} \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \qquad (9.20)$$

which is given by Murthy (1957) and so on.

Empirical Studies. To compare the efficiencies of the above estimators $\hat{y}_1$ to $\hat{y}_{10}$ empirically, the following small population is considered.

unit	u_1	u_2	u_3	u_4
value of p	0.1	0.2	0.3	0.4
value of y	0.5	1.2	2.1	3.2

This population was considered by Yates and Grundy (1953). The sampling scheme consists in selecting 2 units with the specified probabilities without replacement. There are 6 possible samples. The estimates for these samples and the variances of the estimators are given in Table 1.

Table 1. Showing the estimates for all possible samples of 2 units in case of different estimators.

sample	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	$\hat{y}_5$
12	12.0044	4.8514	5.5294	4.3082	8.7669
13	11.3724	5.5844	6.1250	5.0711	8.7487
14	11.1000	6.6021	6.8000	6.2828	8.9500
23	6.8443	6.1714	6.5333	5.8313	6.6888
24	6.2860	7.1891	7.1429	7.1833	6.7144
34	4.7563	7.9221	7.5386	8.3427	6.1474
V(ŷ)	6.5037	0.8281	0.3168	1.5849	1.1123
	$\hat{y}_6$	$\hat{y}_7$	$\hat{y}_8$	$\hat{y}_9$	$\hat{y}_{10}$
12	4.9188	8.4276	8.1563	5.1904	4.5798
13	5.5980	8.4384	8.2218	5.8547	5.3278
14	6.5414	8.8510	8.6914	6.7010	6.4424
23	6.1823	6.5078	6.3378	6.3524	6.0014
24	7.1631	6.7376	6.7346	7.1660	7.1862
31	7.9406	6.3392	6.5495	7.7304	8.1324
V(ŷ)	0.8243	0.8650	0.6625	0.5417	1.5936

Table 2. Showing the different unbiased estimates of the variance of the estimators $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$, $\hat{y}_4$ obtained by estimating y^2 in different ways.

sample	$\hat{v}_1(\hat{y}_1)$	$\hat{v}_2(\hat{y}_1)$	$\hat{v}_3(\hat{y}_1)$	$\hat{v}_4(\hat{y}_1)$
12	106.7469	114.3527	113.6937	114.8395
13	81.3871	93.4566	92.4099	94.4688
14	62.9340	79.0373	78.0471	80.3256
23	3.3472	4.9688	4.2816	5.6236
24	-10.0933	-10.9720	-11.1195	-10.7077
34	-26.7134	-35.1184	-34.1342	-36.2282
$E[\hat{v}(\hat{y}_1)]$	6.4939	6.4956	6.4954	6.4959
sample	$\hat{v}_1(\hat{y}_2)$	$\hat{v}_2(\hat{y}_2)$	$\hat{v}_3(\hat{y}_2)$	$\hat{v}_4(\hat{y}_2)$
12	-13.8226	-6.2168	-6.8758	-5.6800
13	-16.7586	-4.6894	-5.7361	-3.6772
14	-16.6883	-0.5850	-1.5752	0.7033
23	- 5.4110	-3.7894	-4.4766	-3.1346
24	2.0761	1.1974	1ø.0499	1.4617
34	13.4239	5.0189	6.0031	3.9091
$E[\hat{v}(\hat{y}_2)]$	0.8169	0.8187	0.8185	0.8190
sample	$\hat{v}_1(\hat{y}_3)$	$\hat{v}_2(\hat{y}_3)$	$\hat{v}_3(\hat{y}_3)$	$\hat{v}_4(\hat{y}_3)$
12	-6.7844	0.8214	0.1624	1.3582
13	-10.4285	1.6407	0.5940	2.6529
14	-14.0360	2.0673	1.0771	3.3556
23	- 0.8132	0.8084	0.1212	1.4632
24	1.4139	0.5352	0.3877	0.7995
34	7.4947	-0.9103	0.0739	-2.0201
$E[\hat{v}(\hat{y}_3)]$	0.3084	0.3103	0.3100	0.3105
	$\hat{v}_1(\hat{y}_4)$	$\hat{v}_2(\hat{y}_4)$	$\hat{v}_3(\hat{y}_4)$	$\hat{v}_4(\hat{y}_4)$
12	-18.7981	-11.1923	-11.8513	-10.6555
13	-22.2280	-10.1588	-11.2055	- 9.1466
14	-20.8024	- 4.6991	- 5.6893	- 3.4108
23	- 9.4931	- 7.8715	- 8.5587	- 7.2168
24	1.9927	1.1140	0.9665	1.3783
34	20.2648	11.8598	12.8440	10.7500
$E[\hat{v}(\hat{y}_4)]$	1.5735	1.5752	1.5750	1.5755

From Tables (1) and (2), we see that the estimator $\hat{y}_3$ has the least variance and that $\hat{v}_3(\hat{y}_3)$ has the least variability. Incidentally $\hat{v}_3(\hat{y}_3)$ is the only non-negative unbiased variance estimator of all the unbiased variance estimators considered here.

The estimators $\hat{y}_1$ to $\hat{y}_{10}$ have also been studied for the following population, where 2 units are selected with probabilities proportional to x without replacement for estimating the population total y.

unit	u_1	u_2	u_3	u_4
x	2	2	4	6
y	2	6	6	10

This population has been considered by Goodman and Hartley (1958).

Table 3. Showing the estimates for all possible samples of 2 units in case of different estimators.

sample	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	$\hat{y}_5$
12	56.0231	24.1837	27.9992	21.2801	42.0112
13	25.4699	16.1232	17.8181	14.6122	21.6440
14	22.4090	19.5049	19.6000	19.1090	21.0045
23	38.2044	28.2151	30.5452	26.2145	34.3748
24	29.8784	31.5968	30.7995	31.8716	18.3389
34	13.8168	23.5363	22.2963	24.8992	18.0566
$V(\hat{y})$	116.8106	22.3495	22.0736	25.7334	49.7024
sample	$\hat{y}_6$	$\hat{y}_7$	$\hat{y}_8$	$\hat{y}_9$	$\hat{y}_{10}$
12	24.6396	40.1034	38.6516	26.0914	22.7319
13	16.2152	20.7966	20.0410	16.9706	15.3677
14	19.3545	20.9569	20.7590	19.5024	19.3070
23	28.3798	33.2098	32.2094	29.3802	27.2148
24	31.3356	30.7376	30.8750	31.1982	31.7342
34	23.5978	18.6765	19.3580	22.9163	24.2178
$V(\hat{y})$	21.8893	42.9926	37.6330	21.4694	23.5869

From Table 3, we see that though $\hat{Y}_3$ has the least variance of the 4 estimators generated by the technique, some of the combinations of these estimators are more efficient.

Case (ii) : pps and srs of the remaining. Suppose one unit is selected with probability proportional to a given measure of size in the first draw and the other unit is selected with equal probability without replacement in the second draw. Then the estimators corresponding to (9.6) to (9.9) are given by

$$\hat{Y}_1 = \hat{Y}_3 = (y_i + y_j) / (p_i + p_j) \qquad (9.21)$$

$$\hat{Y}_2 = \frac{y_i}{\pi_i} + \frac{y_j}{\pi_j}, \qquad \pi_i = p_i + \frac{(1 - p_i)}{N-1} \qquad (9.22)$$

$$\hat{Y}_4 = \frac{(N-1)}{p_i + p_j} \left[y_i \frac{p_j}{1-p_i} + y_j \frac{p_i}{1-p_j} \right] \qquad (9.23)$$

since

$$P(s) = (p_i + p_j) / (N-1)$$

$$N(s \cdot i) = N-1$$

$$P(s/i1) = 1/(N-1)$$

$$P(s/i2) = p_j / (1-p_i).$$

Of these estimators only (9.21) satisfies the desirable condition mentioned in (5.7), for when we substitute p_i for y_i in (9.21), we get 1.

Case (iii) : srs and pps of the remaining. For the sake of completeness, let us consider the case where one unit is selected with equal probability and then another unit is selected with probability proportional to a given measure of size without replacement.

In this case we get

$$P(s) = \frac{1}{N}\left(\frac{p_i}{1-p_j} + \frac{p_j}{1-p_i}\right)$$

$$N(s \supset i) = (N-1)$$

$$\pi(i) = \pi_i = \sum_{j \neq i} P(s)$$

$$P(s/i1) = p_j / (1-p_i)$$

$$P(s/i2) = \frac{1}{N} \frac{p_i}{1-p_j} \frac{1}{\pi_i - 1/N}$$

The estimators in this case corresponding to those in (9.6) to (9.9) are

$$\hat{Y}_1 = \frac{y_i + y_j}{(N-1)P(s)} \tag{9.24}$$

$$\hat{Y}_2 = \frac{y_i}{\pi_i} + \frac{y_j}{\pi_j} \tag{9.25}$$

$$\hat{Y}_3 = N \frac{y_i p_j(1-p_j) + y_j p_i(1-p_i)}{p_j(1-p_j) + p_i(1-p_i)} \tag{9.26}$$

$$\hat{Y}_4 = \frac{\frac{y_i}{\pi_i - 1/N} p_i(1-p_i) + \frac{y_j}{\pi_j - 1/N} p_j(1-p_j)}{p_i(1-p_i) + p_j(1-p_j)} \tag{9.27}$$

None of the above estimators satisfies the desirable condition (5.7).

In all the above three cases, an unbiased estimator of the population variance σ^2 is given by

$$\frac{(y_i - y_j)^2}{N^2 \pi_{ij}} \tag{9.28}$$

Series of estimators. In the general case the estimators corresponding to those considered above for the particular case of $n = 2$ can be written down. Of course the expressions for the estima-

tors in the general case would be rather complicated in many cases. It is of interest to note that the unordered estimator corresponding to the ordered estimator proposed by Des Raj in sampling with probability proportional to a given measure of size without replacement turns out to be

$$\hat{Y}_3 = \sum_{i=1}^{n} y_i \, P(s/i1) \, / \, P(s).$$

Further such an estimator is obtained by unordering the ordered estimator given by Das when the sampling scheme consists in selecting one unit with probability proportional to a given measure of size in the first draw and then selecting (n-1) units from the remaining units with equal probability without replacement. This matter is discussed in greater detail in chapter 3.

Since the population total Y can be expressed as

$$Y = \frac{1}{\binom{N-1}{m-1}} \sum_m (y_{i1} + y_{i2} + \cdots + y_{im}), \quad (m = 1,2,\ldots n)$$

where $(i1, i2, \ldots, im)$ stands for a specific combination of m units out of N units and $\sum_m$ stands for summation over all combinations of m units out of N units, we get the following unbiased estimators of Y based on a sample selected with probability proportional to size without replacement with the initial probabilities P_i $(i = 1,2,\ldots, N)$:

$$\left.\begin{aligned}
z_1 &= \frac{\sum_1 y_{i_1} P(s/i)}{p(s)} \\
z_2 &= \frac{\sum_2 (y_{i1} + y_{i2}) \, P(s/i_1 i_2)}{(N-1)P(s)} \\
&\vdots \\
z_m &= \frac{\sum_m (y_{i1} + y_{i2} + \cdots, y_{im}) \, P(s/i_1 i_2 \ldots i_m)}{\binom{N-1}{m-1} P(s)} \\
&\vdots \\
z_n &= \frac{y_{i1} + y_{i2} + \cdots + y_{in}}{\binom{N-1}{n-1} P(s)}
\end{aligned}\right\} \qquad (9.29)$$

where $P(s/i_1 i_2, \ldots i_m)$, $(m = 1, 2, \ldots, n)$ stands for the conditional probability of getting the sample 's' given that the units $(i_1, i_2, \ldots)$ have already been selected.

$$
\begin{aligned}
z'_1 &= \sum_1 \frac{y_{i1}}{\pi_{i1}} \\
z'_2 &= \frac{1}{(N-1)} \sum_2 \frac{y_{i1} + y_{i2}}{\pi_{i1\, i2}} \\
&\vdots \\
z'_m &= \frac{1}{\binom{N-1}{m-1}} \sum_m \frac{y_{i1} + y_{i2} + \cdots y_{im}}{\pi_{i1\, i2 \cdots im}} \\
&\vdots \\
z'_n &= \frac{1}{\binom{N-1}{n-1}} \frac{y_{i1} + y_{i2} + \cdots + y_{in}}{\pi_{i1\, i2 \cdots i_n}}
\end{aligned}
\qquad (9.3)
$$

Unlike the previous set of estimators, these estimators are functions of not only the observations and the initial probabilities of selection of the units in the sample, but also of the initial probabilities of all the units in the population and hence are difficult to calculate in practice. It may be observed that the first estimator of this set is that given by Horvitz and Thompson (1952) and that the last estimators of these two sets are the same as that proposed by Midzuno (1950).

In sampling n units with probability proportional to a given measure of size without replacement, some of the unbiased estimators of the population variance σ^2 are given by

$$
\hat{\sigma}_1^2 = \sum_{i=1}^{n} \sum_{j>1} \frac{(y_i - y_j)^2}{N^2 \binom{N-2}{n-2} P(s)} = \frac{\frac{N-1}{N} s^2}{\left(\frac{n}{N}\right) P/s} \qquad (9.31)
$$

$$
\hat{\sigma}_2^2 = \sum_{i=1}^{n} \sum_{j>i} \frac{(y_i - y_j)^2}{N^2 \pi_{ij}} \qquad (9.32)
$$

$$\hat{\sigma}_3^2 = \sum_{i=1}^{n} \sum_{j>1} \frac{(y_i - y_j)^2 P(s/ij1)}{N^2 P(s)} \qquad (9.33)$$

10. SYSTEMATIC SAMPLING

10.1 Equal probability scheme. Two types of simple systematic sampling are considered here - linear and circular. In linear systematic sampling of n units from a finite population of N units, a random start is taken from 1 to k (= N/n) and every kth unit is selected in the sample till the last unit in the population is crossed. In this case there are k possible unordered samples. In circular systematic sampling, the random start is selected from 1 to N and every kth unit is selected proceeding cyclically till n units are selected in the sample. In this case there are N possible samples and one is selected at random. These two procedures are equivalent when N is a multiple of n.

Let us equally distribute the probability of selecting an unordered sample (s) over all possible arrangements of the units in the sample. Then we have in case of linear systematic sampling,

$$P(s) = 1/k$$

$$N(s \ni i) = 1$$

$$\pi(i) = 1/k$$

$$P(s/im) = 1, \ (m = 1, 2, \ldots, n)$$

In this case the different specifications of the generating event considered above lead to the same estimator

$$\hat{Y} = k \sum_{i=1}^{n} y_i \qquad (10.1)$$

The variance of this estimator cannot be unbiasedly estimated in this case, since some pairs of the units are not included in any sample.

In case of circular systematic sampling with equal probability the probability of selection of the unordered sample (s) may be distributed equally over all possible arrangements of the units in the sample. In this case also the different specifications of the generating event lead to one estimator, namely,

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i \qquad (10.2)$$

since

$$P(s) = 1/N$$

$$N(s \ni i) = n$$

$$\pi(i) = n/N$$

$$P(s/im) = 1/n, \ (m = 1, 2, \ldots, n).$$

10.2. Varying probability scheme. Now let us consider the case of probability proportional to size (pps) systematic sampling. Suppose we wish to select a sample of n units such that the probabilities of inclusion of the units in the sample (π_i's) proportional to a given measure of size (X_i, i = 1, 2, ..., N). The procedure of pps systematic sampling consists in cumulating the sizes ($C_i = C_{i-1} + X_i$) (i = 1, 2, ..., N) and then selecting a systematic sample of n units with a random start from 1 to k ($=C_N/n$) and with k as the sampling interval using the cumulative totals. That is the units corresponding to the numbers

$$(R + jk), \ j = 0, 1, 2, \ldots (n-1)$$

are to be selected, R being the random start. The unit U_i is selected in the sample if

$$C_{i-1} < R + jk \leq C_i$$

for some j (= 0, 1, 2, ..., (n-1)). If C_N is not a multiple of n we may select a circular systematic sample.

Here again let us distribute equally the probability of selecting an unordered sample over all possible arrangements of units in the sample. In this we get

$$P(s) = n/X, \quad (X = \sum_{i=1}^{N} X_i)$$

$$N(s \ni i) = X_i$$

$$\pi(i) = n X_i / X$$

$$P(s/il) = 1/X_i$$

In this case also all the above mentioned specifications of the generating event lead to the same estimator

$$\hat{Y} = \frac{X}{n} \sum_{i=1}^{n} \frac{y_i}{x_i} \qquad (10.3)$$

This estimator satisfies the desirable condition (5.7), for when we substitute x_i for y_i in it, we get X.

11. STRATIFIED SAMPLING

Suppose there are k strata and let n_i units be selected from N_i units in the ith stratum (i = 1, 2, ..., k) with equal probability without replacement. In this case we get

$$P(s) = 1 \Big/ \prod_{j=1}^{k} \binom{N_j}{n_j}$$

$$N(s \supset ij) = \binom{N_i - 1}{n_i - 1} \prod_{j \neq i} \binom{N_j}{n_j}$$

$$\pi(ij) = \frac{n_i}{N_i}$$

$$P(s/ijm) = 1 \Big/ \binom{N_i - 1}{n_i - 1} \prod_{j \neq i} \binom{N_j}{n_j} \quad (m = 1, 2, \ldots, n_i).$$

where (ij) stands for the j th unit in the i th stratum.

These specifications of the generating event lead to the same estimator

$$\hat{Y} = \sum_{i=1}^{k} \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij} \qquad (11.1)$$

We could as well have got this estimator considering sampling in the strata separately.

Let us consider the sampling procedure suggested by Murthy, Nanjamma and Sethi (1959). The procedure consists in selecting one unit from the whole population with probability proportional to x (say, U_{ij}) and then selecting $(n_i - 1)$ units from $(N_i - 1)$ units in the ith stratum and n_j units from the N_j units in the jth stratum $(j \neq i)$ with equal probability without replacement. With this sampling scheme we get

$$P(s) = \frac{\sum_{i=1}^{k} N_i \bar{x}_i}{X \prod \binom{N_j}{n_j}}$$

$$N(s \supset ij) = \binom{N_i - 1}{n_i - 1} \prod_{j \neq i} \binom{N_j}{n_j}$$

$$P(s/ijm) = 1 \Big/ \binom{N-1}{n_i - 1} \prod_{j \neq i} \binom{N_j}{n_j}.$$

The unbiased estimator corresponding to the above specifications of the generating event is

$$\frac{\sum_{i=1}^{k} N_i \bar{y}_i}{\sum_{i=1}^{k} N_i \bar{x}_i} X .$$

The estimator involving $\pi(i)$ is not considered as the expression for it turns out to be complicated and hence this is not likely to be very useful in practice.

12. TWO STAGE SAMPLING

Suppose we select n first stage units from N first stage units with equal probability without replacement and from the ith selected first stage units we select n_i second stage units with equal probability without replacement. With this sampling scheme we get

$$P(s) = 1/ \binom{N}{n} \prod_{j=1}^{n} \binom{N_j}{n_j}$$

$$N(s \supset ij) = \binom{N-1}{n-1} \binom{N_i-1}{n_i-1} \prod_{j \neq i}^{n} \binom{N_j}{n_j}$$

$$\pi(ij) = \frac{n}{N} \frac{n_i}{N_i}$$

$$P(s/ij1) = 1/\binom{N-1}{n-1} \binom{N_i-1}{n_i-1} \prod_{j \neq 1} \binom{N_j}{n_j}.$$

All these specifications of the generating event lead to the same estimator

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij} \qquad (12.1)$$

It may be easily verified that for the sampling scheme where one second stage unit is selected from the whole population with probability proportional to a given measure of size (x) (say U_{ij}) and (n_i-1) second

stage units from (N_i-1) second stage units in the ith first stage unit and n_j second stage units from N_j second stage units in the jth selected first stage unit with equal probability without replacement an unbiased estimator is given by

$$\frac{\sum_{i=1}^{n} N_i \bar{y}_i}{\sum_{i=1}^{n} N_i \bar{x}_i} X \qquad (12.2)$$

where $\bar{y}_i$ and $\bar{x}_i$ are sample means for the ith first stage unit, since

$$P(s) = \left(\sum_{i=1}^{n} N_i \bar{y}_i \right) / X \binom{N-1}{n-1} \prod_{j=1}^{n} \binom{N_j}{n_j}$$

$$P(s/ijl) = 1/ \binom{N-1}{n-1} \binom{N_i-1}{n_i-1} \prod_{j \neq i} \binom{N_j}{n_j}$$

In general whatever be the sampling scheme, P(s) and P(s/ijl) may be written in the form

$$P(s) = P_1(s)\, P_{2/1}(s)$$

$$P(s/ijl) = P_1(s/ijl)\, P_{2/1}(s/ijl)$$

where subscripts '1' and '2/1' denote respectively the first stage sampling and the second stage sampling given the selected first stage units. An unbiased estimator is given by

$$\hat{Y} = \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{m_i} y_{ij}\, p_2(s/ijl)}{P_{2/1}(s)} \right] \frac{P_1(s/il)}{P(s)} \qquad (12.3)$$

13. CONCLUSION

The technique of generating unbiased estimators is introduced in section 6 and this technique has been applied to a number of particular situations in the last few sections. This technique systematizes the question of getting unbiased estimators. This is likely to be of much help in deriving unbiased estimators in case of complicated sampling schemes. Another point of interest in this generalized formulation of the problem of unbiased estimation is that it helps in getting unbiased estimators not only for population total, but also for second and higher order moments and other parametric functions which can be expressed as a sum of single valued set functions defined overa class of sets of units in the population.

We have seen that the general estimator obtained by defining the generating event as the occurrence of the set 'a' first in the sample has been found to have some special significance as it covers, as particular cases, the commonly used estimators. Further this general estimator satisfies the desirable condition (5.7) whenever the pps selection is adopted in the first draw.

The treatment in this chapter has been confined to only parametric functions which can be expressed as a sum of single valued set functions defined over a class of sets of elements belonging to the population. In the next chapter, we consider the problem of getting unbiased estimators for parametric functions which can be expressed as non-linear functions of parameters which can be estimated unbiasedly using the technique developed in this chapter.

REFERENCES

1. BASU, D (1958) — On sampling with and without replacement, Sankhya, 20, 287-294.

2. DAS, A.C. (1951) — On two phase sampling and sampling with varying probabilities, Bull. Inter. Stat. Inst. 33(2), 105-112.

3. DES RAJ (1956) — Some estimators in sampling with varying probabilities without replacement, Jour. Amer. Stat. Assn. 51, 269-284.

4. GODAMBE, V. P. (1955) — A unified theory of sampling from finite populations, Jour. Roy. Stat. Soc., 17(B), 269-278.

5. GOODMAN, L. A. (1953) — A simple method for improving some estimators, Ann. Math. Stat. 24, 114-117.

6. GOODMAN, L. A. and HARTLEY, H. O. (1958) — The precision of unbiased ratio type estimators, Jour. Amer. Stat. Assn., 53, 491-508.

7. HANUMANTHA RAO, T.(1960) — An existence theorem in sampling from finite populations (an abstract), Proc. Ind. Sci. Cong., Part III, 35.

8. HORVITZ, D. G. and THOMPSON, D.J. (1952) — A generalization of sampling without replacement from a finite universe, Jour. Amer. Stat. Assn., 47, 663-685.

9. MIDZUNO, H. (1950) — An outline of the theory of sampling systems, Ann. Inst. Stat. Math., 1, 149-156.

10. MURTHY, M. N. (1957) — Ordered and unordered estimators in sampling without replacement, Sankhya, 18, 379-390.

11. MURTHY, M.N. (1960) — A note on probability sampling, Proc. Ind. Sci. Cong., Part III, 34.

12. MURTHY, M.N., NANJAMMA, N.S. and SETHI, V.K. (1959) — Some sampling systems providing unbiased ratio estimators, Sankhya, 21, 292-314.

13. YATES, F. and GRUNDY, P.M. (1953) — Selection without replacement from within strata with probability proportional to size, Jour.Roy. Stat. Soc., B, 15, 253-261.

Chapter 2

ESTIMATION OF BIAS

1. INTRODUCTION

In the last chapter we considered the question of providing unbiased estimator for a certain class of parameters and suggested a technique of generating unbiased estimators for any sample design. The class of parameters was taken to consist of all parameters which can be expressed as a sum of single valued set functions defined over a class of sets of units belonging to the population under consideration. As illustrated earlier, the population total Y and the population variance σ^2, which can be expressed as

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

and

$$\sigma^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j>i} (Y_i - Y_j)^2$$

respectively, are examples of such parameters.

In this chapter, we shall supplement the generalised theory of unbiased estimation given in chapter 1 by giving a procedure of obtaining unbiased (or almost unbiased) estimators for the class of parameters which can be expressed as single valued <u>non-linear</u>

functions of parameters each of which can be expressed as a sum of single valued set functions defined over a class of sets of units belonging to the population. Examples of such parameters are given by ratio of population totals of two characteristics, population standard deviation, correlation coefficient, etc.

The procedure of obtaining unbiased estimator consists in estimating the bias of an estimator, which is taken as the same non-linear function of unbiased estimators of the parameters as the parametric function under consideration, on the basis of independent interpenetrating sub-sample estimates. This procedure is a generalization of the technique used by Murthy and Nanjamma (1959) in estimating the bias of a ratio estimator and by Murthy (1961) in estimating the bias of a product estimator.

The procedure given in this chapter is likely to be of much help in survey practice, since the estimation of relationships between characteristics and between parameters, such as a ratio of population totals of two characteristics are usually of much interest in sample surveys. So far as the procedure has been to take biased but consistent estimators of such parameters and no general procedure was available to estimate the bias in such estimators. In this chapter, it is proposed to fill up this gap in the field of unbiased estimation.

2. PARAMETRIC FUNCTION

Let the parametric function $f(\theta)$ be a single valued non-linear function of the parameters $(\theta_1, \theta_2, \ldots, \theta_k)$, where θ_i $(i=1,2,\ldots, k)$

can be expressed as in (4.1) of chapter 1, namely,

$$\theta_i = \sum_{a_i \in A_i} f_i(a_i) \qquad (2.1)$$

where $f_i(a_i)$ is a single valued set function defined over the class 'A_i' of sets 'a_i' consisting of units belonging to the population X.

Suppose we have defined the sample space 'S' of samples 's' with a suitable probability measure such that it is possible to estimate the parameters $(\theta_1, \theta_2, \ldots, \theta_k)$ unbiasedly using the procedure given in chapter 1. That is, it is assumed that the sample space is so specified that each $a_i \in A_i$ $(i = 1, 2, \ldots, k)$ occurs in at least one 's' and that each 's' contains at least one set 'a_i' in 'A_i' $(i = 1, 2, \ldots, k)$. Then a generalized unbiased estimator of θ_i $(i = 1, 2, \ldots, k)$ is given by

$$t_i = \hat{\theta}_i = \sum_{a_i \subset s} f_i(a_i)\, \phi_i(s, a_i) / P(s) \qquad (2.2)$$

where

$$\sum_{s \supset a_i} \phi_i(s, a_i) = 1.$$

In fact, we can make the above formulation more general by relaxing the assumption that θ_i's $(i = 1, 2, \ldots, k)$ are estimated from the same samples. In other words, θ_i $(i = 1, 2, \ldots, k)$ may be estimated on the basis of the same, overlapping or non-overlapping samples drawn with the same or different sample designs.

Let $(t_1, t_2, \ldots, t_k)$ be unbiased estimators of the parameters $(\theta_1, \theta_2, \ldots, \theta_k)$. Then an estimator of $f(\theta)$ can be taken as $f(t)$ If $f(\theta)$ is a linear function, obviously $f(t)$ will be unbiased for $f(\theta)$. But here we are taking $f(\theta)$ as a non-linear function of $(\theta_1, \theta_2, \ldots, \theta_k)$ and hence $f(t)$ will, in general, be biased for $f(\theta)$.

3. BIAS AND MEAN SQUARE ERROR

In this section approximate expressions for the bias and the mean square error of the estimator $f(t)$ are obtained by using Taylor series symbolically. It may be noted that in statistical practice one is interested not so much in the convergence properties of the infinite series representing a function, but in finding out whether the first few terms of that series will give a good approximation to the function. Because of this, the question of the validity of the application of Taylor series expansion to the case of a finite population estimator will not be considered here. However it will be assumed that the estimator t_i is such that $\left|\frac{t_i - \theta_i}{\theta_i}\right| < 1$, especially for estimators occurring in the denominator of the function $f(t)$ so that the first few terms of the expansion can be expected to give a good approximation to the function. This latter statement has been empirically verified in the case of applying this expansion to a ratio estimator.

If the sample size is fairly large, the assumption $\left|\frac{t_i - \theta_i}{\theta_i}\right| < 1$ will be valid. Let $t_i = \theta_i (1 + e_i)$, $(i = 1,2,\ldots, k)$ and

$$\underline{t} = (t_1, t_2, \ldots, t_k),\ \underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k),\ \underline{e} = (e_1, e_2, \ldots, e_k).$$

Expanding $f(t)$ in a Taylor series about $\underline{t} = \underline{\theta}$ and neglecting terms of degree greater than 2 in e's, we get

$$f(t) = f(\theta) + \sum_{i=1}^{k} \theta_i e_i \left(\frac{\partial f}{\partial t_i}\right)_{t=\theta} + \frac{1}{2}\left[\sum_{i=1}^{k} \theta_i^2 e_i^2 \left(\frac{\partial^2 f}{\partial t_i^2}\right)_{t=\theta} + 2 \sum_{i=1}^{k} \sum_{j>i} \theta_i \theta_j e_i e_j \left(\frac{\partial^2 f}{\partial t_i \partial t_j}\right)_{t=\theta}\right] \quad \ldots \quad (3.1)$$

It may be observed that for certain parameters there will be no terms of degree greater than 2 to neglect. An example of such a parameter is the product $\theta_1 \theta_2$ with the estimator $t_1 t_2$. Taking expected value of $f(t)$ in (3.1), we find that the bias of $f(t)$ correct to the second degree of approximation is given by

$$B[f(t)] = \frac{1}{2}\left[\sum_{i=1}^{k} \left(\frac{\partial^2 f}{\partial t_i^2}\right)_{t=\theta} \mu_2(ii) + 2 \sum_{i=1}^{k} \sum_{j>i} \left(\frac{\partial^2 f}{\partial t_i \partial t_j}\right)_{t=\theta} \mu_2(ij)\right] \quad \ldots \quad (3.2)$$

where $\mu_2(ij) = E(t_i - \theta_i)(t_j - \theta_j),\ (i,j = 1,2,\ldots, k)$.

The mean square error of $f(t)$ to the second degree of approximation is given by

(3.3)

4. BIASES OF TWO ESTIMATORS

Suppose the sample on which the estimate t_i of θ_i $(i = 1, 2, \ldots k)$ is based is selected in the form of n independent interpenetrating sub-samples. Let t_{is} be the unbiased estimate of θ_i based on the sth independent interpenetrating sub-sample $(i = 1, 2, \ldots, k)$, $(s = 1, 2, \ldots, n)$. In this case let us consider the following two estimators T_1 and T_n of $f(\theta)$.

$$T_1 = \frac{1}{n}\sum_{s=1}^{n} f(t_s) \tag{4.1}$$

where $t_s = (t_{1s}, t_{2s}, \ldots, t_{ks})$, $(s = 1, 2, \ldots, n)$, and

$$T_n = f(\bar{t}) \tag{4.2}$$

where $\bar{t} = (\bar{t}_1, \bar{t}_2, \ldots, \bar{t}_k)$, $\bar{t}_i = \frac{1}{n}\sum_{s=1}^{n} t_{is}$, $(i = 1, 2, \ldots, k)$

Applying the result (3.2) to T_n in (4.2) we get

$$B(T_n) = \frac{1}{2}\Big[\sum_{i=1}^{k} \Big) \bar{\mu}_2(ii) + 2\sum_{i=1}^{k}\sum_{j>i} \Big)\bar{\mu}_2(ij)\Big]$$

where $\bar{\mu}_2(ij) = E(\bar{t}_i - \theta_i)(\bar{t}_j - \theta_j) = \frac{1}{n^2}\sum_{s=1}^{n} \mu_{2s}(ij)$

$$\mu_{2s}(ij) = E(t_{is} - \theta_i)t_{js} - \theta_j).$$

That is

$$R_n = B(T_n) = \frac{1}{n^2}\sum_{s=1}^{n} \frac{1}{2}\Big[\sum_{i=1}^{k} \Big) \mu_{2s}(ii) + 2\sum_{i=1}^{k}\sum_{j>i} \Big)$$

$$= \frac{1}{n^2}\sum_{s=1}^{n} B\,[f(t_s)]. \tag{4.3}$$

The bias of the estimator T_1 in (4.1) is given by

$$B_1 = B(T_1) = \frac{1}{n}\sum_{s=1}^{n} B\,[f(t_s)\,] \qquad (4.4)$$

Comparing (4.3) and (4.4) we find that the bias of the estimator T_1 is n times that of the estimator T_n.

5. ESTIMATION OF BIAS

As observed in section 4, comparing the biases of the estimators T_1 and T_n, we get

$$B_1 = n\,B_n \qquad (5.1)$$

Using this result we can derive an unbiased estimator of the bias B_1.

$$E(T_1) = f(\theta) + B_1$$

$$E(T_n) = f(\theta) + B_n$$

Hence $E(T_1 - T_n) = B_1 - B_n = (n-1)\,B_n$.

Thus an unbiased estimator of B_n is given by

$$\hat{B}_n = \frac{T_1 - T_n}{n-1} \qquad (5.2)$$

The variance of the estimator of B_n is given by

$$V(\hat{B}_n) = \frac{V(T_1)}{(n-1)^2}(a^2 - 2a + 1) \qquad (5.3)$$

where $a^2 = V(T_n)/V(T_1)$, and ρ is the correlation coefficient between the estimators T_1 and T_n. For most of the sample designs a^2 and ρ will tend to 1 as the sample size increases and hence the variance of the bias estimator will tend to 0 as sample size increases. It may be observed that an unbiased estimator of the bias of T_1 is given by

$$\hat{B}_1 = \frac{n}{n-1}(T_1 - T_n) \qquad (5.4)$$

6. (ALMOST) UNBIASED ESTIMATOR

Since an unbiased estimator of the bias of the estimator T_n has been obtained in the last section, the estimator T_n can be corrected for its bias, thereby obtaining an unbiased or almost unbiased estimator of $f(\theta)$ according as the third and higher degree terms in 'e' become 0 or not. In the latter case, the estimator is said to be almost unbiased since it is unbiased only to the second degree of approximation. The estimator corrected for its bias is given by

$$T_c = T_n - \hat{B}_n = T_n - \frac{T_n - T_1}{(n-1)} = \frac{nT_n - T_1}{(n-1)} \qquad (6.1)$$

It may be noted that this is the corrected estimator we get, even if we correct the estimator T_1 for its bias.

The variance of the corrected estimator is

$$V(T_c) = \frac{V(T_1)}{(n-1)^2}(n^2a^2 - 2\rho a + 1). \qquad (6.2)$$

The gain in precision in using T_c instead of T_n is given by

$$G(T_c) = \frac{M_n - V(T_c)}{M_n} = 1 - \frac{n^2a^2 - 2n\rho a + 1}{(n-1)^2(a^2 + \rho^2)} \qquad (6.3)$$

where z^2 is the ratio of the square of the bias of T_1 to the variance of T_1. If the sub-sample size is large z^2 will be negligibly small. Neglecting z^2 in the above expression, we find that the gain in precision will be positive if

$$(2n-1)a^2 - 2n\rho a + 1 < 0$$

which will be true if 'a' lies between the roots of the quation

$$(2n - 1)a^2 - 2n\rho a + 1 = 0 \qquad (6.4)$$

For given values of a and ρ, the minimum value of n which makes the corrected estimator more efficient and the value of n which maximisesx the gain are respectively given by

$$\left[\frac{(1 - a^2)}{2a(\rho - a)}\right] + 1 \qquad (6.5)$$

and

$$\frac{(1 - a)}{a(\rho - a)} \qquad (6.6)$$

Table 1 : Showing the minimum and maximum values of $G(T_c)$ and the corresponding values of n for different values of ρ and a ($\rho > a$).

sr. no.	a	ρ	minimum		maximum	
			n	$G(T_c)$	n	$G(T_c)$
1	0.6	0.7	6	0.0089	10	0.0192
2		0.8	3	0.0556	4	0.0988
3		0.9	2	0.0889	3	0.3056
4	0.7	0.8	4	0.0113	7	0.0266
5		0.9	2	0.1020	3	0.1684
6	0.8	0.9	3	0.0469	4	0.0486

(Source : Murthy, M.N. and Nanjamma, N.S. (1959) 'Almost unbiased ratio estimates based on interpenet4ating sub-sample estimates', Sankhya, 21, 381-392).

7. ILLUSTRATIONS

In this particular section, the results derived in the previous sections are applied to some particular cases. The application of this technique of obtaining unbiased estimators to product and ratio estimators are considered in detail in chapters 4 and 5 respectively.

Case (i) : $f(\theta) = \theta^k$. Let t be an unbiased estimator of θ based on any sample design. Then the estimator of $f(\theta)$ is given by

$$f(t) = t^k. \tag{7.1}$$

The bias and mean square error of $f(t)$ correct to the second degree of approximation are given by

$$B[f(t)] = \tfrac{1}{2} k(k-1) C^2 f(\theta) \tag{7.2}$$

$$M[f(t)] = k^2 C^2 [f(\theta)]^2 \tag{7.3}$$

where C^2 is the relative variance of t $[= V(t)/\theta^2]$, since

$$\frac{df}{dt} = k\, t^{k-1} \text{ and } \frac{d^2f}{dt^2} = k(k-1)\, t^{k-2}.$$

The bias relative to the mean square error is

$$\frac{B^2[f(t)]}{M[f(t)]} = \frac{1}{4}(k-1)^2 C^2 \tag{7.4}$$

From (7.2) and (7.4) we see that the bias of $f(t)$ and its contribution to the mean square error both decrease as the sample size increases, since for most sample designs C^2 decreases with increase in sample size.

If t_s ($s = 1, 2, \ldots, n$) are unbiased estimates of θ based on n independent interpenetrating sub-samples, the following two estimators T_1 and T_n of $f(\theta)$ can be considered.

$$T_1 = \frac{1}{n} \sum_{s=1}^{n} t_s^k \tag{7.5}$$

and

$$T_n = \bar{t}^k, \quad (\bar{t} = \frac{1}{n} \sum_{s=1}^{n} t_s) . \tag{7.6}$$

We have seen that the bias of T_1 is n times that of the bias of T_n. Hence an unbiased estimator of the bias of T_n is given by

$$B(T_n) = \frac{\sum_{s=1}^{n} t_s^k - n\bar{t}^k}{n(n-1)} \tag{7.7}$$

and the corrected estimator is given by

$$T_c = \frac{n^2\bar{t}^k - \sum_{s=1}^{n} t_s^k}{n(n-1)} . \tag{7.8}$$

It may be noted that the expression for bias and the corrected estimator will be completely unbiased if k in $f(\theta)$ is 2.

Case (ii) : Correlation Coefficient ( ). The correlation coefficient between two characteristics x and y is

$$\frac{\mathrm{Cov}(x,y)}{V(x)V(y)} . \tag{7.9}$$

In this case the parametric function is of the form

$$f(\theta) = \frac{\theta_1}{\theta_2 \theta_3} \tag{7.10}$$

and the estimator is given by

$$f(t) = \frac{t_1}{t_2\, t_3} \qquad (7.11)$$

where t_1, t_2 and t_3 are unbiased estimators of θ_1, θ_2 and θ_3 respectively. The bias and mean square error of f(t) correct to the second degree approximation are given by

$$B[f(t)] = \frac{f(\theta)}{\theta}\,[3(v_{22} + v_{33}) - 4(v_{12} + v_{13}) + 2v_{23}] \qquad (7.12)$$

and

$$M[f(t)] = \frac{[f(\theta)]^2}{4}\,[4v_{11} + (v_{22} + v_{33}) - 4(v_{12} + v_{13}) + 2v_{23}] \qquad (7.13)$$

...

where

$$v_{ij} = \frac{E(t_i - \theta_i)(t_j - \theta_j)}{\theta_i\, \theta_j}.$$

Let t_{is} (i = 1, 2, 3) be unbiased estimates based on the sth independent interpenetrating sub-sample (s = 1, 2, ..., n). Then using the two estimators

$$T_1 = \frac{1}{n} \sum_{s=1}^{n} \frac{t_{is}}{t_{2s}\, t_{3s}} \qquad (7.14)$$

$$T_n = \frac{t_1}{t_2\, t_3} \qquad (7.15)$$

we get the following corrected estimator of

$$T_c = \frac{n\, T_n - T_1}{(n-1)}\,. \qquad (7.16)$$

Case (iii) : Regression Estimator. Let y and x be unbiased estimators of the population totals Y and X respectively and let b be consistent estimator of the regression coefficient obtained

by taking unbiased estimators of the covariance between x and y and the variance of x. The regression estimator is $\bar{y} = y + b(X-x)$. (7.17)

The estimator in this case is of the form

$$f(t) = t_1 + \frac{t_2}{t_3}(X - t_4) \qquad (7.18)$$

The bias and mean square error of this estimator correct to the second degree of approximation are given by

$$B[f(t)] = \beta \times (v_{34} - v_{24}) \qquad (7.19)$$

and

$$M[f(t)] = V(y) - 2\beta\, Cov(x,y) + \beta^2 V(x) \qquad (7.20)$$

By defining the two estimators T_1 and T_n on the basis of n independent interpenetrating sub-sample estimates, we get the corrected estimator as

$$T_c = \frac{nT_n - T_1}{n-1}$$

Case (iv) : Skewness.

$$(\beta_2 = \frac{\mu_4}{\mu_2^2})$$

The parametric function is of the form

$$f(\theta) = \frac{\theta_1}{\theta_2^2}$$

and an estimator of $f(\theta)$ is given by

$$f(t) = \frac{t_1}{t_2^2}$$

where t_1 and t_2 are unbiased estimators of θ_1 and θ_2 respectively. The bias and the mean square error of f(t) correct to the second

degree of approximation are given by

$$B\left[f(t) \right] = \beta_2 \; (3 \; \; \;)$$

and

$$M\left[f(t) \right] = \beta_2^2 \; (\; \; \;)$$

where

Defining suitably the two estimators T_1 and T_n based on n independent interpenetrating sub-sample estimates, we get the corrected estimator, as before, as

$$\hat{\beta}_{2.} = \frac{n\,T_n - T_1}{n - 1}$$

8. ESTIMATION OF BIAS (GENERAL CASE)

Suppose $f(\theta)$ is the parametric function of the parameters $(\theta_1, \theta_2, \ldots, \theta_k)$ and $f(t)$ is an estimator of $f(\theta)$ based on the estimators $(t_1, t_2, \ldots, t_k)$ which are unbiased for the parameters $(\theta_1, \theta_2, \ldots, \theta_k)$. Let $t_i = \theta_i + h_i$, $i = 1, 2, \ldots, k$. Applying Taylor series expansion to $f(t)$ about $t = \theta$ symbolically and neglecting terms of degree greater than p in h_i's, we get

$$f(t) = f(\theta) + \sum_{j=1}^{p} \frac{1}{j!} \sum_{i_1, i_2, \ldots i_j} (h_{i_1} h_{i_2} \ldots h_{i_j}) \left(\frac{d^j f}{dt_{i_1} dt_{i_2} \ldots dt_{i_j}} \right)_{t=\theta} \qquad (8.1)$$

Taking the expected value of (8.1), we get the bias of $f(t)$ as

$$B[f(t)] = \sum_{j=2}^{p} \frac{1}{j!} \sum_{i_1} \sum_{i_2,\ldots,i_j} \mu_j(i_1, i_2, \ldots i_j) \left[\frac{d^j f}{dt_{i_1} dt_{i_2} \ldots dt_{i_j}} \right]_{\ldots} \qquad (8.2)$$

Suppose t_{is} is an unbiased estimate of θ_i based on the sth independent interpenetrating sub-sample (i=1,2,..., k, s = 1,2,...n). Let us consider the following p estimators of $f(\theta)$

$$T_m = \frac{1}{\binom{n}{m}} \sum f(\bar{t}_{(m)}), (m = 1, 2, \ldots, p-1, n) \qquad (8.3)$$

where

$$\bar{t}(m) = (\bar{t}_1(m), \bar{t}_2(m), \ldots, \bar{t}_k(m))$$

$\bar{t}_i(m)$ being the mean of the estimate t_i based on a combination of m sub-samples taken from the n independent interpenetrating sub-samples and $\sum$ denotes summation over all combinations of m sub-samples formed out of n sub-samples.

The bias of T_m to the pth degree of approximation is given by

$$B_m = B(T_m) = \frac{1}{\binom{n}{m}} \sum B[f(\bar{t}(m))]$$

$$= \frac{1}{\binom{n}{m}} \sum E \left[\sum_{j=2}^{p} \frac{1}{j!} \sum_{i_r} (\bar{h}_{i_1} \bar{h}_{i_2} \ldots \bar{h}_{i_j}) \right.$$

$$\left. (\qquad) \right]_{\ldots} \qquad (8.4)$$

where $\bar{h}_{i_r} = \frac{1}{m} \sum_{s=1}^{m} h_{i_r s}$. After simplification the bias of T_m may be expressed in the form

$$B_m = \sum_{j=2}^{p} \frac{A_j}{m^{j-1}} \qquad (m = 1,2, \ldots, p-1,n) \qquad (8.5)$$

where A_j is a function of the jth order moments and product moments of the estimators $(t_1, t_2, \ldots, t_k)$ and of terms of the form

$$\left[\frac{d^r f}{dt_{i_1} \, dt_{i_2} \ldots dt_{i_r}} \right]_{t = \theta}$$

From (8.5) we see that in the series of estimators T_m, $B(T_{m+1}) < B(T_m)$

Since $E(T_m) = f(\theta) + B_m$,

we get $$E(T_1 - T_m) = B_1 - B_m = \sum_{j=2}^{p} \left(1 - \frac{1}{m^{j-1}}\right) A_j \qquad (8.6)$$

Let $D_m = (T_1 - T_m)$. The equation (8.6) can be written as

$$E(D) = A \qquad (8.7)$$

where $D = (D_2, D_3, \ldots, D_{p-1}, D_n)$

$$A = (A_2, A_3, \ldots, A_{p-1}, A_p)$$

$$= 1 - \frac{1}{2} \quad 1 - \frac{1}{3} \ldots 1 - \frac{1}{p-1} \quad 1 - \frac{1}{n-1}$$

$$1 - \frac{1}{2^2} \quad 1 - \frac{1}{3^2} \quad \ldots \quad 1 - \frac{1}{(p-1)^2} \quad 1 - \frac{1}{(n-1)^2}$$

$$\vdots \qquad \vdots \qquad \qquad \vdots \qquad \vdots$$

$$1 - \frac{1}{2^{p-1}} \quad 1 - \frac{1}{3^{p-1}} \ldots 1 - \frac{1}{(p-1)^{p-1}} \quad 1 - \frac{1}{(n-1)^{p-1}}$$

It may be noted that in (8.7) we have (p-1) equations in (p-1) unknowns. It may be observed that we are considering p estimators

since there are (p-1) A's and f(θ) to be estimated. Solving (8.7) for A we get

$$A = E(D)^{-1} \tag{8.8}$$

Taking B as $B = (B_2, B_3, \ldots, B_{p-1}, B_n)$, we get

$$B = A\,(e + \;) \tag{8.9}$$

where 'e' is a (p-1, p-1) matrix whose elements are all equal to 1. Substituting in (8.9) the solution for A obtained in (8.8) we get unbiased estimators of the biases of the estimators, namely,

$$B = D^{-1} e - D.$$

That is,

$$B_m = \sum_j D_j S_{j-1} - D_m, \quad (j = 2, 3, \ldots, p-1, n),$$

$$(m = 2, 3, \ldots, p-1, m) \tag{8.10}$$

$S_{n-1} = S_{p-1}$ where S_j is the sum of the elements in the jth row of

Particular cases

(i) p = 2.

This is the case considered earlier. In this case, the following 2 estimators of f(θ) may be considered.

$$T_1 = \frac{1}{n} \sum_{j=1}^{n} f(t_s)$$

$$T_n = f(\bar{t}).$$

$$B_m = A/m, \quad (m = 1, n).$$

Since $= (1 - \frac{1}{n})$, $A = \frac{n}{n-1}$, $S_1 = \frac{n}{n-1}$, we get

$$B_n = \frac{n}{n-1}(T_1 - T_n) - (T_1 - T_n)$$

$$= (T_1 - T_n) / (n-1).$$

(ii) $p = 3$.

Let us consider the following three estimators of $f(\theta)$

$$T_1 = \frac{1}{n}\sum_{s=1}^{n} f(t_s) \tag{8.11}$$

$$T_2 = \frac{2}{n(n-1)} \sum f(\bar{t}(2)) \tag{8.12}$$

$$T_n = f(t). \tag{8.13}$$

$$B_n = \frac{A_2}{n} + \frac{A_3}{n^2}, \quad (n = 1, 2, n \tag{8.14}$$

Since $1 - \frac{1}{2}$ $1 - \frac{1}{n}$

$1 - \frac{1}{2^2}$ $1 - \frac{1}{n^2}$

and

$$= - \frac{4n^2}{(n-1)(n-2)} \quad \frac{n^2-1}{n^2} - \frac{n-1}{n}$$

$$-\frac{3}{4} \qquad \frac{1}{2}$$

we get after simplification

$$B_1 = \frac{n-2}{n-1} T_1 + \frac{4}{n-2} T_2 - \frac{n^2}{(n-1)(n-2)} T_n \tag{8.15}$$

$$B_2 = \frac{-1}{n-1} T_1 + \frac{n+2}{n-2} T_2 - \frac{n^2}{(n-1)(n-2)} T_n \tag{8.16}$$

$$B_n = -\frac{1}{n-1} T_1 + \frac{4}{n-2} T_2 - \frac{(3n-2)}{(n-1)(n-2)} T_n \tag{8.17}$$

REFERENCES

1. MURTHY, M. N. and NANJAMMA, N. S. (1959) Almost unbiased ratio estimates based on interpenetrating sub-sample estimates, Sankhya, 21, 381-392.

2. MURTHY, M. N. (1961) Unbiased product estimators based on interpenetrating sub-sample estimates, submitted for publication in Jour. Roy. Stat. Soc.

Chapter 3

ORDERED AND UNORDERED ESTIMATORS

1. INTRODUCTION

In this chapter a technique is given to improve upon estimators which are based on the order of selection of units in sampling without replacement. This technique has already been mentioned in section 3 of chapter 1. In sampling without replacement, Das (1951) and Des Raj (1956) have given certain estimators which take into account the order of selection of the units in the sample. Such estimators may be termed 'ordered' estimators. Roy Chowdhury (1956) has shown that one of the 'ordered' estimators suggested by Des Raj is more efficient than the estimator for sampling with replacement. Another advantage of this estimator is that it admits of a non-negative variance estimator.

In this chapter it is shown that corresponding to any biased or unbiased 'ordered' estimator there exists an 'unordered' estimator which is more efficient than the former for any convex risk function. By 'unordered' estimator is meant an estimator which ignores the order of selection of the units in the sample. The technique of 'unordering' an ordered estimator is illustrated. This method is applied to the set of ordered estimators of population total given

by Das and Des Raj and also to the unbiased variance estimators considered by them.

It is shown that the technique of unordering preserves the desirable properties of one of the ordered estimators mentioned above. In sampling the first unit with varying probability and the rest with equal probability without replacement, unordering of Das' ordered estimator yields the familiar unbiased ratio estimator. It is of interest to note that the general forms of the unordered estimators considered here can be generated by the technique of generating estimators introduced in chapter 1.

2. UNORDERED ESTIMATOR

In sampling n units without replacement from a finite population of N units, there will be $\binom{N}{n}$ unordered samples (s). Corresponding to any unordered sample (s) of size n units, there will be n! ordered samples (si). Let x_{si} $\left[s = 1, 2, \ldots \binom{N}{n}\ i = 1, 2, \ldots, N\ (=n!)\right]$ be an estimator of population parameter θ based on the ordered sample (si). Consider a scheme of selection in which the probability of selecting the ordered sample (si) is p_{si}. Then the probability p_s of getting the unordered sample (s) is the sum of the probabilities of getting the ordered samples corresponding to (s)

That is, $$p_s = \sum_{i=1}^{N} p_{si}.$$

THEOREM 1 : If $\hat{\theta}_o = x_{si}$ and $\hat{\theta}_u = \sum_{i=1}^{M} x_{si}\, p'_{si}$ (where $p'_{si} = p_{si}/p_s$)

are estimators of the population parameter θ, then

$$\text{(i) } E(\hat{\theta}_u) = E(\hat{\theta}_o)$$

and $$\text{(ii) } V(\hat{\theta}_u) \leq V(\hat{\theta}_o)$$

where E and V stand for expectation and variance respectively.

Proof : $$E(\hat{\theta}_u) = \sum_{s=1}^{\binom{N}{n}} \hat{\theta}_u p_s = \sum_{s=1}^{\binom{N}{n}} \sum_{i=1}^{M} x_{si}\, p_{si} = E(\hat{\theta}_o)$$

The variances of the estimators $\hat{\theta}_o$ and $\hat{\theta}_u$ are given by

$$V(\hat{\theta}_o) = \sum_{s=1}^{\binom{N}{n}} \sum_{i=1}^{M} x_{si}^2\, p_{si} - \left(\sum_{s=1}^{\binom{N}{n}} \sum_{i=1}^{M} x_{si}\, p_{si} \right)^2 \quad (2.1)$$

$$V(\hat{\theta}_u) = \sum_{s=1}^{\binom{N}{n}} \left(\sum_{i=1}^{M} x_{si}\, p'_{si} \right)^2 p_s - \left(\sum_{s=1}^{\binom{N}{n}} \sum_{i=1}^{M} x_{si}\, p_{si} \right)^2 \quad (2.2)$$

Therefore, $$V(\hat{\theta}_o) - V(\hat{\theta}_u) = \sum_{s=1}^{\binom{N}{n}} \sum_{i=1}^{M} \left(x_{si} - \sum_{i=1}^{M} x_{si}\, p'_{si} \right)^2 p_{si} . \quad (2.3)$$

This shows that the variance of the unordered estimator $\hat{\theta}_u$ is less than or equal to that of the ordered estimator $\hat{\theta}_o$.

Corollary:

(i) If $\hat{V}_o(\hat{\theta}_o) = v_{si}$ is an ordered estimator of $V(\hat{\theta}_o)$, then $V_u(\hat{\theta}_o)$, an unordered estimator of $V(\hat{\theta}_o)$, which has a lesser mean square error than $\hat{V}_o(\hat{\theta}_o)$ is given by

$$\hat{V}_u(\hat{\theta}_o) = \sum_{i=1}^{M} v_{si}\, p'_{si} . \quad (3.4)$$

(ii) The mean square error of $\hat{\theta}_U$ is less than or equal to that of $\hat{\theta}_c$.

(iii) An estimator $\hat{V}_U(\hat{\theta}_U)$ of the variance of $\hat{\theta}_U$ is given by

$$\hat{V}_U(\hat{\theta}_U) = \hat{\theta}_U^2 - \sum_{i=1}^{M} (x_{si}^2 - v_{si})p'_{si} . \qquad (2.5)$$

(iv) The 2k-th moment of $\hat{\theta}_U$ is less than or equal to that θ.

That is, $E[\hat{\theta}_U - E(\theta_U)]^{2k} \leq E[\hat{\theta}_c - E(\theta_c)]^{2k}$. (2.6)

It may be noted that estimator based on $m(m = 2, \ldots, M-1)$ ordered samples drawn from the M ordered samples corresponding to the unordered sample (s) with probability proportional to the sum of their probabilities will be more efficient than the ordered estimator $\hat{\theta}_o$.

3. UNORDERING OF DES RAJ's ESTIMATORS

Let a sample of size n be drawn from a finite population of N units with varying probabilities without replacement. Suppose the probabilities of selection at the first draw are

$$p_j, (j = 1, 2, \ldots, N),\ p_j > 0,\ \sum_{j=1}^{N} p_j = 1 . \qquad (3.1)$$

The scheme of selection of a unit at a particular draw depends on the units already drawn in the sample and not on the order in which they were drawn. For instance, the probabilities of selection at the third draw given that the k-th and l-th units have already been chosen in the first two draws will be given by

$$\frac{p_j}{1 - p_k - p_1} \quad (j \neq k \neq 1).$$

Let $(y_1, y_2, \ldots, y_n)$ and $(p_1, p_2, \ldots p_n)$ be the values of the units arranged in the order of selection in the sample drawn according to the above scheme and their respective ini ial probabilities. Des Raj (1958) considered three sets of ordered estimators one of which was originally given by Das (1951). One of the sets of estimators given by Des Raj is given by

$$x_{si1} = \frac{y_1}{p_1}$$

$$x_{si2} = y_1 + \frac{y_2}{p_2}(1 - p_1)$$

$$x_{sin} = y_1 + y_2 + \cdots + y_{n-1} + \left(\frac{y_n}{p_n}\right)(1 - p_1 - p_2 - \cdots - p_{n-1})$$

Each of the above estimators is unbiased for the population total, and, therefore,

$$\bar{x}_{si} = \frac{1}{n}\sum_{j=1}^{n} x_{sij} \tag{3.3}$$

is also so. By making use of the fact that x_{sij} and $x_{sij'}$, $(j \neq j')$ are uncorrelated, Des Raj was able to get a non-negative estimator of the variance of $\bar{x}_{si}$ which is given by

$$\hat{V}_e(\bar{x}_{si}) = v_{si} = \frac{1}{n(n-1)}\sum_{j=1}^{n}(x_{sij} - \bar{x}_{si})^2 \tag{3.4}$$

By applying the earlier theorem to $\bar{x}_{si}$ and v_{si} we get more efficient estimators of the population total and the variance of

$\bar{x}_{si}$ respectively.

THEOREM 2 : Unordering of the ordered estimator

$$\theta_o = \sum_{j=1}^{n} c_j x_{sij}, \sum_{j=1}^{n} c_j = 1, \; x_{sij} = y_1 + y_2 + \cdots + y_{j-1} + \frac{y_j}{p_j}(1 - p_1 - \cdots p_{j-1}), \qquad (3.5)$$

yields an unordered estimator which is independent of the set

$$c_j \; , \sum_{j=1}^{n} c_j = 1 \; , \text{ namely,}$$

$$\frac{\sum_{i=1}^{n} y_i P(s/i)}{P(s)} \qquad (3.6)$$

where $P(s/i)$ is the conditional probability of getting the unordered sample (s) given that i-th unit has been selected at the first draw and p(s) is the unconditional probability of getting (s).

Proof : Let $P(s/i_1, i_2 \ldots i_j)$ denote the conditional probability of getting the unordered sample (s) given that i_1-th, i_2-th , ..., i_j th units have been selected in the first j draws.

The coefficient of y_i in the estimator got by unordering θ_o in the usual way (Theorem 1) is given by $\frac{1}{P(s)}$ times

The coefficient of e_1 in the above expression is $P(s/i)$. The theorem will be proved if we show that the coefficient of e_{j+1} is equal to that of e_j in the expression (3.7).

The first (j-1) terms in both the coefficients are the same. The j-th and (j+1)th terms in the coefficient of e_{j+1} reduce to the j-th term in the coefficient of e_j because of the equality

$$P(s/i\ i_1\ i_2 \dots i_{j-1}) = \sum_{\substack{i_j \\ i_j \neq i_{j-1} \neq \dots \neq i_1 \neq i}} \frac{p_{ij}}{(1 - p_i - p_{i_1} - \dots p_{i_{j-1}})} \times P(s/i\ i_1\ i_2 \dots, i_j).$$

Therefore $$\theta_u = x_s \frac{\sum_{i=1}^{n} y_i P(s/i)}{P(s)}.$$

The estimator given in (3.6) is a special case of the general estimator obtained in chaper 1 by defining the generating event as the occurrence of the i-th unit in the first draw.

Since the unbiased estimators x_{sij} and $x_{sij'}$ $(j \neq j')$ of the population total Y are uncorrelated, we get

$$\sum_{s=1}^{\binom{N}{n}} \sum_{i=1}^{N} (x_{sij}\ x_{sij'})\ p_{si} = Y^2.$$

As this is true for all the $\binom{n}{2}$ pairs of the n estimators.

$$\sum_{s=1}^{N} \sum_{i=1}^{M} \left(\sum_{j>j'}^{n} x_{sij}\, x_{sij'} \right) p_{si} = \frac{n}{2} Y^2$$

Hence an unbiased estimator of Y^2 is given by

$$(\hat{Y}^2) = \frac{1}{\frac{n}{2}} \sum_{i=1}^{M} \left(\sum_{j>j'} x_{sij}\, x_{sij'} \right) p'_{si} \qquad (3.8)$$

From this it follows that an unbiased estimator of the variance of $\bar{x}_s$ is

$$\hat{V}_U(\bar{x}_s) = \hat{x}_s^2 - \frac{1}{\frac{n}{2}} \sum_{i=1}^{M} \left(\sum_{j>j'}^{n} x_{sij}\, x_{sij'} \right) p'_{si}$$

$$= \frac{1}{[P(s)]^2} \left[\sum_{i=1}^{n} P(s/i)\, P(s/i) - P(s) \right] y_i^2 + 2 \sum_{i>i'}^{n} P(s/i)P(s/i') - P(s)P(s/ii') \; y_i y_{i'} \Big]$$

The variance estimator may be expressed as

$$\hat{V}_u(\bar{x}_s) = \frac{1}{[P(s)]^2} \sum_{i=1}^{n} \sum_{j>i} \left\{ P(s)\, P(s/ij) - P(s/i)\, P(s/j) \right\} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \qquad (3.9)$$

for $\sum_{i=1}^{n} P(s/i) \left\{ P(s/i) - P(s) \right\} y_i^2 = \sum_{i=1}^{n} \sum_{j>i} \left\{ P(s)\, P(s/ij) - P(s/i)\, P(s/j) \right\} p_i p_j \left(\frac{y_i^2}{p_i^2} + \frac{y_j^2}{p_j^2} \right)$.

It is rather difficult to show that (3.9) is non-negative in general. We shall show that this is non-negative for n = 2 and 3 and it is expected that this will be so in general. For n = 2, the variance estimator will be non-negative if

$$P(s)\,P(s/ij) - P(s/i)\,P(s/j) \leq 0$$

if $$\frac{p_i p_j (2 - p_i - p_j)}{(1 - p_i)(1 - p_j)} - \frac{p_j}{(1-p_i)} \cdot \frac{p_i}{(1-p_j)} \leq 0$$

if $(1 - p_i - p_j) \leq 0$

which is true. Similarly in case of n = 3, it can be shown that the variance estimator is non-negative for in that case $P(s)\,P(s/ij) - P(s/i)\,P(s/j)$ reduces to

$$(1 - p_k)(1 - p_i - p_j)$$

$$(1 - p_i - p_j - p_k) + (1 - p_i)(1 - p_j)(1 - p_i - p_j)$$

which is non-negative.

The following particular cases of equations (3.6) and (3.9) will now be considered.

(i) Simple random sampling without replacement. In this case the estimators $\bar{x}_{si}$, $\bar{x}_s$ and their variances are given by

$$\bar{x}_{si} = \frac{1}{n} \sum_{j=1}^{n} (N + n + 1 - 2j)y_j \qquad (3.10)$$

$$\bar{x}_s = \frac{N}{n} \sum_{j=1}^{n} y_j \qquad (3.11)$$

$$V(\bar{x}_{si}) = \frac{\sigma^2}{n} N(N-n) + \frac{n^2-1}{3} \qquad (3.12)$$

and

$$V(\bar{x}_s) = \frac{\sigma^2}{n} N(N-n) \qquad (3.13)$$

where

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2 \qquad (3.14)$$

Therefore

$$V(\bar{x}_{si}) - V(\bar{x}_s) = \frac{n^2 - 1}{3}.$$

Comparison of the above expressions for the variances shows that $V(\bar{x}_s) \leq V(\bar{x}_{si})$. This is otherwise obvious also as $N\bar{y}$ is known to be the ~~first~~ unbiased linear estimator of Y. It is interesting to note that the variance estimator of $\bar{x}_s$ given in equation (3.9) reduces to the estimator commonly used, namely,

$$V_0(\bar{x}_s) = N(N-n)\frac{s^2}{n} \qquad (3.15)$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

It may be pointed out that the divergence of Des Raj's estimator from the best unbiased linear estimator led to a search for a more efficient estimator.

(ii) Sampling of two units with varying probabilities without replacement.

This case is of importance as in actual practice one will, in general, be choosing two units from each stratum in stratified sampling. In this case $\bar{x}_{si}$ and $\bar{x}_s$ are given by

$$\bar{x}_{si} = \tfrac{1}{2}(1+p_1)\frac{y_1}{p_1} + (1-p_1)\frac{y_2}{p_2} \qquad (3.16)$$

and

$$\bar{x}_s = \frac{1}{2-p_1-p_2}(1-p_2)\frac{y_1}{p_1} + (1-p_1)\frac{y_2}{p_2} \qquad (3.17)$$

The sampling variances of these two estimators are given by

$$V(\bar{x}_{si}) = \frac{1}{4} \sum_{s=1}^{\binom{N}{2}} p_1 p_2 (2-p_1-p_2) \left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \qquad (3.18)$$

and

$$V(\bar{x}_s) = \sum_{s=1}^{\binom{N}{2}} p_1 p_2 \frac{(1-p_1-p_2)}{(2-p_1-p_2)} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \qquad (3.19)$$

Therefore,

$$V(\bar{x}_{si}) - V(\bar{x}_s) = \frac{1}{2} \sum_{s=1}^{\binom{N}{2}} p_1 p_2 \frac{(p_1+p_2)^2}{(2-p_1-p_2)} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \qquad (3.20)$$

In the case of simple random sampling without replacement the expression (3.18) above becomes

$$V(\bar{x}_{si}) - V(\bar{x}_s) = \frac{\sigma^2}{2},$$

which is a particular case of the expression (3.14).

The estimator of variance of $\bar{x}_{si}$ given by Des Raj is,

$$v_o(\bar{x}_{si}) = v_{si} = \frac{1}{4}(1-p_1)^2 \left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \qquad (3.21)$$

By applying the earlier theorem to this v_{si}, we get a more efficient estimator of the variance, namely,

$$v_o(\bar{x}_{si}) = \frac{1}{4}(1-p_1)(1-p_2)\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \qquad (3.22)$$

and this also is non-negative. Substituting the relevant values in equation (3.7) we get an unbiased estimator of the variance of $\bar{x}_s$, namely

$$v_o(\bar{x}_s) = \frac{(1-p_1)(1-p_2)(1-p_1-p_2)}{(2-p_1-p_2)^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \qquad (3.23)$$

which is always non-negative.

4. UNORDERING OF DAS' ESTIMATORS

With the notation adopted in section 3, the set of estimators proposed by Das is given by,

$$x'_{si1} = \left(\frac{y_1}{p_1}\right)$$

$$x'_{si2} = \frac{1}{(N-1)} \frac{(1-p_1)}{p_1} \left(\frac{y_2}{p_2}\right)$$

$$x'_{sin} = \frac{(1-p_1)(1-p_1-p_2) \ldots (1-p_1-p_2- \ldots - p_{n-1})}{(N-1)(N-2) \ldots (N-n+1) p_1 p_2 \cdots p_{n-1}} \left(\frac{y_n}{p_n}\right)$$

... (4.1)

Each of the above estimators is unbiased for the population total Y and so is their mean

$$\bar{x}'_{si} = \frac{1}{n} \sum_{j=1}^{n} x'_{sij} . \qquad (4.2)$$

An unbiased estimator of the variance of $\bar{x}'_{si}$ is

$$\hat{V}_o(\bar{x}'_{si}) = v_{si} = x'^2_{si} - \frac{1}{n} \sum_{j=1}^{n} (x'_{sij}\, y_j) + \frac{N-1}{\binom{n}{2}} \sum_{j>j=1}^{n} (x'_{sij}\, y_j) . \qquad (4.3)$$

THEOREM 3 : Unordering of the ordered estimators

$$\bar{x}'_{si} = \frac{1}{n} \sum_{j=1}^{n} x'_{sij}$$

yields the unordered estimator

where $P(s/i_1, i_2 \ldots i_j)$ is the conditional probability of getting the sample 's' given that the units $(i_1, i_2, \ldots, i_j)$ have occurred in the first j draws and $\sum_j$ stands for summation over all combinations of j units out of n units.

Proof. In unordering $\bar{x}'_{s_i}$ the coefficient of y_i occurring at the j-th draw is

$$\frac{1}{n} \frac{1}{\binom{N-1}{j-1}} \sum' \frac{P(s/i_1, i_2, \ldots, i_j)}{P(s)}$$

where $\sum'$ denotes summation over all possible ordered samples where y_i occurs at the jth place. Simplifying this, we get

$$\frac{1}{n} \frac{1}{\binom{N-1}{j-1}} \sum' \frac{P(s/i_1, i_2, \ldots, i_j)}{P(s)}$$

where $\sum'$ stands for summation over $(i_1, i_2, \ldots, i_{j-1})$. Hence we get

$$\frac{1}{n} \sum_{j=1}^{n} \frac{1}{\binom{N-1}{j-1}} \sum' \frac{P(s/i_1, i_2, \ldots, i_j)}{P(s)} y_i .$$

It may be noted that the unordered estimator in this case is the mean of all the estimators obtained by defining the set 'a' in different ways. The generating event considered here is of the a type E_{a1} given in chapter 1.

In sampling the first unit with varying probability and the remaining (n-1) units from (N-1) units with equal probability without

replacement, the estimator $\bar{x}'_{si}$ becomes

$$\bar{x}'_{si} = \frac{1}{n} \frac{\sum_{j=1}^{n} y_j}{p_1} \qquad (4.5)$$

Applying the Theorem 1 to this $\bar{x}'_{si}$ we get the unordered estimator

$$\bar{x}'_s = \frac{\sum_{j=1}^{n} y_j}{\sum_{j=1}^{n} p_j} \qquad (4.6)$$

This shows that the above unbiased ratio estimator is more efficient than he estimator given by Das. Lahiri (1951) and Midzuno (1952) have given sampling procedures which lead to the above estimator.

Substituting the relevant values in equation (2.5) we get an unbiased variance estimator of $\bar{x}'_s$ namely,

$$\hat{V}(\bar{x}'_s) = \bar{x}'_s - (N-1) \frac{(\sum_{j=1}^{n} y_j)^2 - (N-n)(\sum_{j=1}^{n} y_j^2)}{(n-1)(\sum_{j=1}^{n} p_j)} \qquad (4.7)$$

5. UNORDERING OF DES RAJ's SECOND SET

Another set of ordered estimators is obtained using the unconditional probability of a unit occurring in a particular draw. The estimators are given by

$$x''_{si1} = \frac{y_1}{p_1(1)}$$

$$x''_{si2} = \frac{y_2}{p_2(2)}$$

$$x''_{sij} = \frac{y_j}{p_j(j)}$$

$$x''_{sin} = \frac{y_n}{p_n(n)}$$

where $p_j(i)$ is the unconditional probability of getting y_i at the j-th draw. The combined estimator may be taken as

$$\bar{x}^{11}_{si} = \frac{1}{n} \sum_{j=1}^{n} x^{11}_{sij} \qquad (5.1)$$

THEOREM 4 : Unordering of $\bar{x}^{11}_{si}$ gives rise to the unordered estimator

$$\bar{x}^{11}_{s} = \frac{1}{n} \sum_{j=1}^{n} \frac{\sum_{i=1}^{n} y_i P(s/ij)}{P(s)} \qquad (5.2)$$

where P(s/ij) is the conditional probability of getting the sample s given that the i-th unit is selected at the j-th draw. The coefficient of y_i occurring at the j-th draw is given by

where P(s, ij) is the probability of getting the sample 's' with y_i at the j-th place. Since y_i can occur in any one of the n draws, we get

$$\frac{1}{n} \sum_{j=1}^{n} P(s/i_j)$$

as the coefficient of y_i in the unordered estimator? Hence we get

$$\bar{x}^{11}_{s} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\sum y_i \, P(s/i_j)}{P(s)} \right)$$

It may be noted that in this case the unordered estimator is the mean of the estimators obtained by defining the generating event as the occurrence of the unit with value y_i in the j-th draw (j=1,2,...n).

6. NUMERICAL EXAMPLES

To study the relative performance of the ordered and unordered estimators, the following population given by Yates and Grundy (1953) will be considered.

unit	p	y	y/p
1	.1	0.5	5
2	.2	1.2	6
3	.3	2.1	7
4	.4	3.2	8
total	1.0	7.0	

This was deliberately chosen by them as being more extreme than will normally be encountered in practice. The object is to estimate the population total by selecting two units. Two schemes of selection will be considered.

Case (i): first unit with varying probability and second unit with equal probability without replacement,

Case (ii) : both the units with varying probabilities without replacement.

For the purpose of comparison, the estimator and the variance estimator proposed by Horvitz and Thompson (1952) together with the variance estimator given by Yates and Grundy are also considered.

For the sake of convenience, the expressions for the different estimators given in subsequent tables are given below. Suppose (y_1, y_2) is the ordered sample drawn from a finite population of N units.

Case (i)

$$\bar{x}_{si} = \tfrac{1}{2}\left[(1 + p_1)\left(\frac{y_1}{p_1}\right) + (N-1)y_2\right] \tag{6.1}$$

$$\hat{V}_0(\bar{x}_{si}) = v_{si} = \frac{1}{4}\left[(1-p_1)\left(\frac{y_1}{p_1}\right) - (N-1)y_2\right]^2 \tag{6.2}$$

$$\hat{V}_0(\bar{x}_{si}) = \frac{v_{s1}\,p_1 + v_{s2}\,p_2}{p_1+p_2} \tag{6.3}$$

$$\bar{x}_s = \frac{1}{2(p_1+p_2)}\left[\{1+p_1+(N-1)p_2\}\,y_1 + \{1 + p_2 + (N-1)p_1\}\,y_2\right] \tag{6.4}$$

$$\hat{V}_0(\bar{x}_s) = \hat{V}_0(\bar{x}_{si}) - \frac{(\bar{x}_{s1} - \bar{x}_s)^2\,p_1 + (\bar{x}_{s2} - \bar{x}_s)^2 p_2}{p_1 + p_2} \tag{6.5}$$

$$\bar{x}'_{si} = \frac{y_1 + y_2}{2p_1} \tag{6.6}$$

$$\hat{V}_0(\bar{x}'_{si}) = v'_{si} = \frac{1}{4}\left[\left(\frac{1}{p_1} - 1\right)y_1 + (N-1)y_2\right]^2 \tag{6.7}$$

$$\hat{V}_0(\bar{x}'_s) = \frac{v'_{s1}\,p_1 + v'_{s2}\,p_2}{p_1+p_2} \tag{6.8}$$

$$\bar{x}'_s = \frac{1}{2(p_1+p_2)}\left[\left\{\frac{1}{p_1}+p_1+(N-1)p_2\right\}y_1 + \{1 + p_2 + (N-1)p_1\}\,y_2\right] \tag{6.9}$$

$$\hat{V}_0(\bar{x}'_s) = \hat{V}_0(\bar{x}'_{si}) - \frac{(\bar{x}'_{s1} - \bar{x}'_s)^2\,p_1 + (\bar{x}'_{s2} - \bar{x}'_s)^2 p_2}{p_1 + p_2} \tag{6.10}$$

and

$$\hat{V}_0(\bar{x}'_s) = \frac{1}{4}\left[\left\{\frac{1}{p_1}+p_1+(N-1)p_2\right\}y_1 + \{1 + p_2 + (N-1)p_1\}\,y_2\right]^2 \tag{6.11}$$

$$V_{HT}(y_{HT}) = (1-\pi_1)\left(\frac{y_1^2}{\pi_1^2}\right) + (1-\pi_2)\left(\frac{y_2^2}{\pi_2^2}\right) + 2\,\frac{\pi_{12} - \pi_1\pi_2}{\pi_{12}}\,\frac{y_1\,y_2}{\pi_1\,\pi_2}$$

where

$$\pi_{12} = \frac{p_1 + p_2}{(N-1)} \tag{6.12}$$

$$V_{YG}(y_{HT}) = \frac{\pi_1\pi_2 - \pi_{12}}{\pi_{12}}\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right) \tag{6.13}$$

Case (ii)

$$x_{si} = \tfrac{1}{2}(1 + p_1)\left(\frac{y_1}{p_1}\right) + (1 - p_1)\left(\frac{y_2}{p_2}\right)$$

$$V_0(x_{si}) = \tfrac{1}{4}(1-p_1)^2\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2$$

$$V_U(x_{si}) = \tfrac{1}{4}(1-p_1)(1-p_2)\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2$$

$$x_s = \frac{1}{2-p_1-p_2}\,(1-p_2)\left(\frac{y_1}{p_1}\right) + (1 - p_1)\left(\frac{y_2}{p_2}\right)$$

$$V_U(x_s) = \frac{(1p_1)(1-p_2)(1-p_1-p_2)}{(2-p_1-p_2)^2}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2$$

$$x_{si} = \tfrac{1}{2}\left(\frac{y_1}{p_1}\right) + \frac{1}{N-1}\left(\frac{1-p_1}{p_1}\right)\left(\frac{y_2}{p_2}\right) \tag{6.14}$$

$$V_0(x_{si}) = v'_{si} = x_{si}^2 - \tfrac{1}{2}\left(\frac{y_1^2}{p_1}\right) + \frac{1}{N-1}\,\frac{(1-p_1)}{p_1}\left(\frac{y_2^2}{p_2}\right) + \frac{2(1-p_1)}{p_1}\,\frac{y_1y_2}{p_2} \tag{6.15}$$

$$V_U(x_s) = \frac{1}{2-p_1-p_2}(1-p_2)v'_{s1} + (1-p_1)v'_{s2} \tag{6.16}$$

$$x_s = \frac{1}{2-p_1-p_2}(1-p_2)x_{s1} + (1-p_1)x_{s2} \tag{6.17}$$

Table 1 – Unbiased Estimates of error variance case (i)

sample	$\hat{V}_o(\bar{x}_{\bar{s}})$	$\hat{V}_U(\bar{x}_{si})$	$\hat{V}_U(\bar{x}_s)$	$\hat{V}_o(\bar{x}'_{si})$	$\hat{V}_U(\bar{x}'_{si})$	$\hat{V}_u(\bar{x}'_s)$	$\hat{V}_{HT}(\hat{y}_{HT})$	$\hat{V}_{YG}(\hat{y}_{HT})$
12	.20	1.88	1.87	45.80	18.40	14.48	-1.11	1.51
13	.81	2.37	2.30	114.20	28.93	14.85	2.27	4.33
14	6.50	3.48	3.25	241.80	45.38	14.58	6.45	7.34
21	2.72	1.88	1.87	4.84	18.49	14.48	-1.11	1.51
23	.56	.48	.44	15.64	3.44	1.62	.77	.92
24	5.76	2.16	1.94	34.20	2.63	-4.09	3.93	3.05
31	2.89	2.37	2.30	.51	28.93	14.85	2.27	4.33
32	.42	.48	.44	-4.70	3.44	1.62	.77	.92
34	5.52	2.69	2.60	-13.59	-20.01	-21.20	3.01	.72
41	2.72	3.48	3.25	- 3.72	45.38	14.58	6.45	7.34
42	.36	2.16	1.94	-13.15	2.63	- 4.09	3.93	3.05
43	.56	2.69	2.60	-24.82	-20.01	-21.20	3.01	.72
true error variance	2.223	2.223	2.103	9.701	9.701	0.363	2.884	2.884
variance of estimated error variance	1.9912	.8357	.7543	2583.03	491.72	194.33	4.7317	5.5115

Table 2 – Unbiased estimates of error variance case (ii)

sample	$\hat{V}_0(\bar{x}_{si})$	$\hat{V}_U(\bar{x}_{si})$	$\hat{V}_U(\bar{x}_s)$	$\hat{V}_0(\bar{x}'_{si})$	$\hat{V}_U(\bar{x}'_{si})$	$\hat{V}_U(\bar{x}'_s)$	$\hat{V}_{HT}(\hat{y}_{HT})$	$\hat{V}_{YG}(y_{HT})$
12	.20	.18	.17	93.20	49.60	42.95	-6.20	.41
13	.81	.63	.59	114.20	48.17	34.12	-4.70	1.52
14	1.82	1.22	1.08	134.60	47.92	27.39	- .58	2.79
21	.16	.18	.17	10.84	49.60	42.95	-6.20	.41
23	.16	.14	.12	11.78	2.55	1.71	-3.79	.36
24	.64	.48	.39	10.38	-3.07	-5.03	1.21	1.08
31	.49	.63	.59	-3.18	48.17	34.12	-4.70	1.52
32	.12	.14	.12	-5.52	2.55	1.71	-3.79	.36
34	.12	.10	.07	-12.80	-15.07	-15.25	5.02	.18
41	.81	1.22	1.08	- 9.86	47.92	27.39	-.58	2.79
42	.36	.48	.39	-13.15	-3.07	-5.03	1.21	1.08
43	.09	.10	.07	-17.01	-15.07	-15.25	5.02	.18
true error variance	0.365	0.365	0.312	5.435	5.435	1.107	0.823	0.823
variance of estimated error variance	0.1575	0.1236	0.1004	1542.49	606.18	350.73	14.8692	.6811

$$V_U(\bar{x}_s) = V_U(x_{si}) - \frac{1}{2-p_1-p_2}\,(1-p_2)(x_{s1} - \bar{x}_s)^2 + (1-p_1)(x_{s2} - \bar{x}_s)^2 \qquad (6.18)$$

$$y_{HT} = \left(\frac{y_1}{\pi_1}\right) + \left(\frac{y_2}{\pi_2}\right) \quad \text{where } \pi_1 = p_1\left[1 + \sum_{j=1}^{N} \left(\frac{p_j}{1-p_j}\right)\right]$$

and
$$\pi_2 = p_2\left[1 + \sum_{j=2}^{N} \left(\frac{p_j}{1-p_j}\right)\right] \qquad (6.19)$$

$$V_{HT}(y_{HT}) = (1-\pi_1)\left(\frac{y_1^2}{\pi_1^2}\right) + (1-\pi_2)\left(\frac{y_2^2}{\pi_2^2}\right) + \frac{\pi_{12} - \pi_1\pi_2}{\pi_{12}}\left(\frac{y_1}{\pi_1}\right)\left(\frac{y_2}{\pi_2}\right)$$

where

$$\pi_{12} = \frac{p_1 p_2 (2 - p_1 - p_2)}{(1-p_1)(1-p_2)}$$

$$V_{YG}(y_{HT}) = \frac{\pi_1\pi_2 - \pi_{12}}{\pi_{12}}\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2. \qquad (6.20)$$

The results given in the above table show that for this population unordering of Das' estimators in the above two cases, yields estimators which are much more efficient than the corresponding ordered estimators. Of the three unordered unbiased estimators of the population total, namely , x_s, $\bar{x}_s$ and y_{HT}, x_s' in case (i) and $\overline{x_s}$ in case (ii) have the least variance. It may be noted that in both the cases the estimator obtained by defining the generating event as the occurrence of a unit in the first draw has turned out to be the most efficient. It can also be seen that it is possible to improve substantially on the ordered variance estimators by unordering them.

REFERENCES

1. DAS, A. C. (1951) : Two phase sampling and sampling with varying probabilities. Bull. Int. Stat. Inst., 33, 105-112

2. DES RAJ (1956) : Some estimators in sampling with varying probabilities wi hout replacement. J. Amer. Stat. Ass., 51, 274, 269-284.

3. HORVITZ, D.G. and THOMPSON, D. I. (1952) : A generalisation of sampling without replacement from a finite universe. J. Amer. Stat. Ass., 47, 663-85.

4. LAHIRI, D.B. (1951) A method of sample selection providing unbiased ratio estimates. Bull. Int. Stat. Inst., 33, 133-140.

5. MIDZUNO, H. (1952) On the sampling system with probability proportional to the sum of sizes. Ann. Inst. Stat. Math., (Japan), 3, 99-108.

6. ROY CHOWDHURY, D.K. (1956) Selection with varying probabilities (part of thesis submitted for the Associateship of the Indian Statistical Institute).

7. YATES, F. and GRUNDY, P. M. (1953) Selection without replacement from within strata with probability propor ional to size. J. Roy. Stat. Soc., B, 15, 253-61.

Chapter 4

RATIO METHOD OF ESTIMATION

1. INTRODUCTION

In practice estimation of the ratio of two population parameters is of considerable importance. For instance, in socio economic surveys, one may be interested to estimate such ratios as income or expenditure per household and per person, proportion of unemployed persons, proportion of expenditure/on different items. Estimation of yield rate in a crop survey and of input output ratio in an industrial survey are of importance. In estimating such ratios, the commonly used procedure has been to take the ratio of unbiased estimators of the numerator and the denominator of the population ratio as an estimator. Further ratio method of estimation is also used in practice to improve conventional estimators with the use of supplementary information on a related characteristic.

A very satisfactory treatment of the question of bias and mean square error of ratio estimators is not yet available. However, in recent years the question of bias of a ratio estimator has received considerable attention. Hartley and Ross (1954) suggested a ratio type estimator which is unbiased, when the value of the parameter occurring in the denominator is known. In this chapter we shall generalize this ratio type estimator

to any sample design using the independent interpenetrating sub-sample estimates.

Murthy and Nanjamma (1959) have given a technique of estimating the bias of a ratio estimator any given degree of ~~ex~~ approximation using independent interpenetrating sub-sample estimates. The estimate of bias may be used to correct the estimator for its bias to any given degree of approximation, thereby obtaining an 'almost' unbiased estimator. It may be mentioned that this technique is a particular case of the generalized technique of obtaining (almost) unbiased estimators for non-linear functions of parameters considered in chapter 2. In this chapter, we shall apply the technique of getting unbiased estimators developed in ~~the last~~ chapter to the question of ratio estimators and derive the results obtained by Murthy and Nanjamma (1959).

Murthy, Nanjamma and Sethi (1959) have suggested simple modifications of many of the sampling schemes commonly adopted in practice, namely, equal probability sampling, varying probability sampling, stratified and multi-stage sampling, which, while retaining the form of the usual ratio estimators, make them unbiased. In this chapter, the efficiencies of the unbiased ratio estimator and of the biased ratio estimator have been compared from the point of view of mean square error. In other words, the variance of the unbiased estimator is compared with the mean square error of the biased ratio estimator. It is shown that in large samples the two estimators are equally efficient to the second

degree of approximation and that the unbiased estimator is more efficient than the corresponding biased ratio estimator to the fourth degree of approximation in large samples, if the estimators of the numerator and the denominator of the ratio are distributed in the bivariate normal form for the original sampling scheme.

2. BIAS AND MEAN SQUARE ERROR.

Let y and x be unbiased estimates of the population totals y and X based on any sample design. Then the estimator

$$\hat{R} = \frac{y}{x} \tag{2.1}$$

of the population ratio $R \left(= \frac{Y}{X} \right)$, is consistent but biased. Let $y = Y(1 + e)$ and $x = X(1 + e')$. Assuming that [illegible] and applying the Taylor series expansion symbolically to $\hat{R}$ at $(y = Y, x = X)$, we get

$$\hat{R} = R[\, 1 + (e - e') + (e'^2 - ee') + (e\,e'^2 - e'^3) + \dots \,] \tag{2.2}$$

From (2.2) it can be shown that the bias of the estimator to the second degree approximation is given by

$$B(\hat{R}) = R \left([illegible] \right) \tag{2.3}$$

and that the mean square error correct to the fourth degree approximation is given by

$$M(\hat{R}) = R^2 \left[\left([illegible] \right) + 2 \left([illegible] \right) + 3 \left([illegible] \right) \right] \tag{2.4}$$

where $\mu_{ij} = \dfrac{E(\bar{x} - \bar{X})^2 (\bar{y} - \bar{Y})^j}{\bar{X}^i \bar{Y}^j}$

It may be noted that the assumption $\left| \dfrac{\bar{x} - \bar{X}}{\bar{X}} \right| < 1$ is likely to be valid if the sample size is fairly large.

3. COMPARISON OF TWO RATIO ESTIMATORS.

Let (y_s, x_s) be unbiased estimates of the population totals Y and X from the sth independent interpenetrating sub-sample ($s = 1, 2, \ldots, n$). The following two estimators can be taken to estimate R,

$$\text{(i)}\ R_1 = \frac{1}{n} \sum_{s=1}^{n} \frac{y_s}{x_s} \tag{3.1}$$

$$\text{(ii)}\ R_n = \frac{y_1 + y_2 + \ldots + y_n}{x_1 + x_2 + \ldots + x_n} \tag{3.2}$$

From the results of chapter 2, it can be easily seen that the bias of the estimator R_1 is n times that of the estimator R_n. That is

$$B_1 = n B_n \tag{3.3}$$

We now compare the mean square errors of R_1 and R_n to the fourth degree of approximation, assuming that the sub-sample sizes are the same (as is the case generally) so that

$$B\left(\frac{y_i}{x_i}\right) = B$$

$$M\left(\frac{y_i}{x_i}\right) = M \text{ for all } i.$$

By applying (2.4) to R_n and simplifying, we obtain

$$M(R_n) = M_n = \frac{R^2}{n}\Big[\Big(\qquad\Big) + \frac{2}{n}\Big(2\qquad\Big)$$

$$+ \frac{3}{n^2}\Big(\qquad\Big)$$

$$+ \frac{3(n-1)}{n^2}\Big(3\qquad\Big)\Big]$$

$$= \frac{M}{n} - \frac{n-1}{n^2} A \tag{3.4}$$

where $A = R^2\Big[\,2\Big(2\qquad\Big) + \frac{3n+1}{n}\Big(\qquad\Big)$

$$- \frac{3}{n}\Big(3\qquad\Big)\Big]$$

From (3.4) and (3.5), we get

$$M_1 = M_n + \frac{n-1}{n^2} A + \frac{n-1}{n} B^2 \tag{3.5}$$

Comparison of M_1 and M_n is difficult in general? If it is assumed that x and y are distributed in the bivariate normal form, we get

$$B = RC_x\,(c_x \text{ —— })$$

$$M = R^2[(c^2_y - 2\,c_x c_y + c^2_x)(1 + 3c^2_x) + 6c^2_x(c_x - c_y)^2]$$

where $c^2_x =$ 20, $c^2_y =$ 02 and is the correlation coefficient

between x and y. Further $A = R^2 c^2_x [(c^2_y - 2 c_x c_y + c^2_x) +$

$$2(c_x - c_y)^2 > 0.$$

$\therefore$ The mean square error of R_1 is greater than that of R_n. Thus R_n is better than R_1 from the considerations of both bias and mean square error.

4. ESTIMATION OF BIAS

Proceeding exactly in the same way as in chapter 2, we get

$$E(R_1) = R + B_1$$

$$E(R_n) = R + B_n$$

$$E(R_1 - R_n) = B_1 - B_n.$$

But $B_n = nB_1$,

$$\therefore E(R_1 - R_n) = (n-1) B_n.$$

$\therefore$ An unbiased estimator of the bias B_n is given by

$$\hat{B}_n = \frac{R_1 - R_n}{n-1} \qquad (4.1)$$

The corrected estimator in this case is

$$R_c = R_n - \hat{B}_n = \frac{nR_n - R_1}{(n-1)} \qquad (4.2)$$

This estimator may be considered as 'almost' unbiased since it is unbiased only to the second degree of approximation. The conditions under which this will be more efficient than R_n will be the same as those derived in chapter 2.

5. BIAS UPTO 3RD DEGREE APPROXIMATION

Let ($\bar{x}_s, \bar{y}_s$) be unbiased estimates of Y and X based on the sth independent interpenetrating sub-sample (s = 1, 2, ... n) If it is required to estimate the bias of the ratio estimator upto third degree approximation, the following three estimators may be considered.

$$R_1 = \frac{1}{n} \sum_{s=1}^{n} \frac{y_s}{x_s}$$

$$R_2 = \frac{1}{\binom{n}{2}} \sum_2 \frac{y_i + y_j}{x_i + x_j}$$

$$R_n = \frac{y_1 + y_2 + \ldots + y_n}{x_1 + x_2 + \ldots x_n}$$

where $\sum_2$ denotes summation over combinations of 2 sub-samples taken from n sub-samples. Applying the theory developed in chapter 2, we get the unbiased estimators of biases of R_1, R_2 and R_n as

$$B_1 = \frac{n-2}{n-1} R_1 + \frac{4}{n-2} R_2 - \frac{n^2}{(n-1)(n-2)} R_n$$

$$B_2 = -\frac{1}{n-1} R_1 + \frac{n+2}{n-2} R_2 - \frac{n^2}{(n-1)(n-2)}$$

$$B_n = -\frac{1}{n-1} R_1 + \frac{4}{n-2} R_2 - \frac{3n-2}{(n-1)(n-2)} R_n$$

Hence the almost unbiased estimator in this case is

$$R_c = \frac{1}{n-1} R_1 - \frac{4}{n-2} R_2 + \frac{n^2}{(n-1)(n-2)} R_n.$$

6. AN ILLUSTRATION (I)

The technique of making a ratio estimator unbiased using independent interpenetrating sub-sample estimates has been applied to the estimates of yield rates of cereal crops. For this study the estimates of crop production and crop acreage given in Report Number 38 of the National Sample Survey have been used. The estimates R_1, R_n and R_c have been obtained. The number of sub-samples is 2 in this case.

Table 1 - Showing the values of the yield rate estimators R_1, R_2, R_c for different cereal crops by zones

(tons/acre)

zone	no. of sample villages*	R_1	R_2	R_c	R_1	R_2	R_c
(0)	(1)	(2)	(3)	(4)	(2)	(3)	(4)
		rice			jowar		
North India	445	0.3917	0.3926	0.3935	0.1672	0.1610	0.1548
Central India	692	0.2451	0.2451	0.2451	0.2746	0.2771	0.2796
East India	721	0.3220	0.3211	0.3202	0.3586	0.3056	0.2526
South India	529	0.5163	0.5164	0.5165	0.2816	0.2819	0.2822
West India	676	0.4045	0.4041	0.4037	0.2978	0.2975	0.2972
All India	3063	0.3562	0.3559	0.3556	0.2852	0.2852	0.2852
		bajra			ragi		
North India	445	0.0633	0.0644	0.0655	-	-	-
Central India	692	0.2418	0.2406	0.2394	0.1666	0.5000	0.8334
East India	721	0.3004	0.2294	0.1584	0.4303	0.4462	0.4621
South India	529	0.1745	0.1753	0.1761	0.4570	0.4514	0.4458
West India	676	0.1664	0.1657	0.1650	0.2972	0.2996	0.3020
All India	3063	0.1278	0.1279	0.1280	0.3758	0.3754	0.3750
		maize			wheat		
North India	445	0.6042	0.6041	0.6040	0.3978	0.3978	0.3964
Central India	692	0.3326	0.3283	0.3240	0.2582	0.2579	0.2576
East India	721	0.3861	0.3976	0.4091	0.1944	0.1943	0.1942
South India	529	0.5144	0.5132	0.5120	0.1930	0.2400	0.2870
West India	676	0.5367	0.5379	0.5391	0.2356	0.2361	0.2366
All India	3063	0.4776	0.4760	0.4744	0.2930	0.2929	0.2928
		barley			seven cereals		
North India	445	0.2538	0.2548	0.2558	0.2420	0.2418	0.2416
Central India	692	0.2318	0.2352	0.2386	0.2580	0.2580	0.2580
East India	721	0.2204	0.2238	0.2272	0.3174	0.3166	0.3158
South India	529	0.2696	0.2647	0.2598	0.2870	0.2870	0.2870
West India	676	0.2696	0.2647	0.2598	0.2870	0.2870	0.2870
All India	3063	0.2379	0.2383	0.2387	0.2910	0.2909	0.2908

* Crop cutting in all seasons and land utilization in autumn season only in one-third of the sample villages.

7. RATIO TYPE ESTIMATOR

In case of ratio method of estimaton for estimating a population total using the data on a suitable supplementary variate, Hartley and Ross (1954) have obtained a ratio-type of estimator which is unbiased in case of simple random sampling. Suppose (y_i, x_i) are the values of the variates y and x for the ith unit in the sample selected with equal probability without replacement (i = 1, 2, ..., n). The bias of the estimator

$$\hat{Y} = \bar{r} X, \quad (\bar{r} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}) \tag{7.1}$$

can be expressed as

$$B(\hat{Y}) = - N \operatorname{cov} \left(\frac{y}{x}, x \right) \tag{7.2}$$

Since cov $\left(\frac{y}{x}, x \right)$ can be unbiasedly estimated, the estimator given in (7.1) can be corrected for its bias and the corrected estimator is given by

$$Y' = \bar{r} X + \frac{N-1}{n-1} n(\bar{y} - \bar{r} \bar{x}). \tag{7.3}$$

If the sampling is with replacement the factor (N-1) is to be replaced by N.

Goodman and Hartley (1958) have shown that for large samples, this unbiased ratio-type estimator is more efficient than the usual combined ratio estimator $\left(\frac{\bar{y}}{\bar{x}} X \right)$ if the slope of the population regression line of y on x is closer to $\frac{1}{N} \sum_{i=1}^{N} \frac{Y_i}{X_i}$ than to $\frac{Y}{X}$. However this condition is not in general satisfied because one would ordinarily use

a ratio estimator only when the regression coefficient is expected to be near Y/X. Hence the above condition is rather restrictive and the proposed unbiased ratio estimator may be less efficient than the usual combined ratio estimator in large samples.

It may be mentioned that this technique of getting an unbiased ratio estimator is applicable only if the value of the denominator of the population ratio is known, which is the case when one is using the ratio method of estimation for estimating the population mean or total using a suitable supplementary variate. But in case of estimating a population ratio where usually the value of the denominator is not known, it is not possible to use this technique.

The above method can easily be generalized to the case where y_i and x_i $(i = 1, 2, \ldots, m)$ are unbiased estimators of the population totals Y and X based on m interpenetrating sub-samples of the same size selected according to any specified sampling design. In this case an unbiased ratio-type estimator of Y is given by

$$Y' = \bar{r} X + \frac{m}{m-1} (\bar{y} - \bar{r} \bar{x}) \qquad (7.4)$$

where $\bar{r} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_i}{x_i}$ and $\bar{y}$ and $\bar{x}$ are means of the sub-sample estimates of Y and X respectively. It can be easily shown with the usual assumptions that this estimator and the ordinary ratio estimator are equally efficient in large samples. This is not so for the estimator (7.3).

8. BIASED RATIO ESTIMATOR

Suppose y and x are unbiased estimators of the population parameters Y and X based on any probability sampling design. Then an estimator of the ratio R(=Y/X) is given by

$$R_o = y/x \tag{8.1}$$

where the subscript 'o' stands for the original sampling scheme. This estimator is biased. Its bias and mean square error can be obtained as follows writing $y = Y(1 + e)$ and $x = X(1+e')$.

$$E_o(R_o) = R\, E\left(\frac{1+e}{1+e'} - 1\right) \tag{8.2}$$

$$M_o(R_o) = R^2\, E\left(\frac{1+e}{1+e'} - 1\right)^2 \tag{8.3}$$

where E_o and M_o denote the expected value and mean square error for the original sampling scheme.

Under the assumption that $|e'| < 1$, the bias and mean square error can be derived in the following manner. This assumption means that for all samples, the value of the estimator x lies between 0 and 2X which is usually the case if the sample size is fairly large.

$$\begin{aligned} B(R_o) &= R\, E_o(e-e')(1-e' + e'^2 - e'^3 + \dots) \\ &= R\, E_o\left[(e - e') + (e'^2 - ee') + (ee'^2 - e'^3) + \dots\right] \\ &= R\left[(v_{20} - v_{11}) + (v_{21} - v_{30}) + (v_{40} - v_{31}) + \dots\right] \end{aligned} \tag{8.4}$$

and

$$\begin{aligned} M_o(R_o) &= R^2\, E_o(e - e')^2(1-2e' + 3e'^2 - 4e'^3 + \dots) \\ &= R^2\, E_o\left[e^2 - 2ee'\,\Delta e'^2) - 2(e^2e' - 2ee'^2 + e'^3) + \dots\right. \\ &= R^2\left[(v_{20} - 2v_{11} + v_{02}) - 2(v_{12} - 2v_{21} + v_{03})\right. \\ &\quad \left. + 3(v_{22} - 2v_{31} + v_{40}) \dots\right] \end{aligned} \tag{8.5}$$

where

$$\nu_{ij} = \frac{1}{\bar{x}^i \bar{y}^j} E_o (x - \bar{x})^i (y - \bar{y})^j.$$

If the sample size is fairly large, the relative moments of order greater than to 2 may be neglected.

9. UNBIASED RATIO ESTIMATOR

If p is the probability of selecting a given sample in the original sampling scheme, the estimator

$$R_m = \frac{\bar{y}}{\bar{x}} \tag{9.1}$$

whose form is exactly the same as that of R_o given in (2.1), can be made unbiased for the population ratio by modifying the original sampling scheme such that a given sample is selected with probability proportional to xp. For in that case the expected value of the ratio estimator is

$$E_m(R_m) = \frac{\bar{y}}{\bar{x}} \frac{xp}{\sum xp} = \frac{\bar{Y}}{\bar{X}}$$

where the subscript m stands for the modified sampling scheme and $\sum$ stands for summation over all possible samples. As mentioned earlier, Murthy, Nanjamma and Sathi (1959) have given sample modifications in the original selection prodcedure which make the probability of selecting the sample proportional to xp.

The variance of the estimator R_m is given by

$$V_m(R_m) = E_m(R_m)^2 - R^2$$

$$= \sum_m \left(\frac{\bar{y}}{\bar{x}}\right)^2 \frac{xp}{X} - R^2$$

$$= \frac{1}{X} E_o \left(\frac{\bar{y}^2}{\bar{x}}\right) - R^2$$

Now writing $y = Y(1 + e)$ and $x = X(1+e')$ as before we get

$$V_m(R_m) = R^2 E_o \left[\frac{(1+e)^2}{(1+e')} - 1 \right].$$

Under the assumption that $e' < 1$, we get

$$V_m(R_m) = R^2 E_o \left[(2e - e') + (e^2 - 2ee' + e'^2) - (e^2e' - 2ee'^2 + e'^3) + (e^2e'^2 - 2ee'^3 + e'^4) - \ldots \right]$$

$$R^2 \left[(v_{02} - 2v_{11} + v_{20}) - (v_{12} - 2v_{21} + v_{30}) + (v_{22} - 2v_{31} + v_{40}) - \ldots \right] \qquad (9.2)$$

where v_{ij} is as defined in (8.5). If the sample size is fairly large, the relative moments of order greater than 2 may be neglected.

10. COMPARISON OF RATIO ESTIMATORS.

Comparing the mean square error of R_o given in (8.5) and the variance of R_m given in (9.2) we see that the two estimators are equally efficient to the second degree of approximation if the terms involving moments of order greater than 2 can be neglected. This will be true if the sample size is fairly large. In fact, for with replacement sampling schemes, this would mean neglecting terms involving $(1/n^2)$ and higher powers of $(1/n)$, where n is the sample size.

Subtracting (8.5) from (9.2) we get to any specified degree of approximation

$$V_m(R_m) - M_o(R_o) = R^2 \left[(v_{12} - 2v_{21} + v_{30}) - 2(v_{22} - 2v_{31} + v_{40}) \ldots \right] \qquad (10.1)$$

In general it is difficult to compare the efficiencies of these two esti-

mators. Assuming that the estimators y and x are distributed in the bivariate normal form for the original sampling scheme, we get correct upto the fourth degree of approximation

$$V_m(R_m) - M_o(R_o) = R^2[(1+2/^{\alpha 2})\, C_x^2\, \chi_y^2 - 6\, C_y C_x^3 + 3\, C_{x-}^4 \,] \quad (10.2)$$

where is the correlation coefficient between the estimators x and y and C_x and C_y are the coefficients of variation of the estimators x and y respectively. Simplifying (10.2) we get

$$V_m(R_m) - M_o(R_o) = -\, 2R^2 C^2{}_x \,[\, C^2{}_y\, (1 - \,) + 3(C_x - C_y)^2 \,] \quad (10.3)$$

which shows that

$$V_m(R_m) < M_o(R_o) \quad (10.4)$$

Thus we see that the unbiased ratio estimator is more efficient than the corresponding biased ratio estimator if moments of order greater than 4 can be neglected.

11. AN ILLUSTRATION (II).

The efficiencies of the two ratio estimators, compared in section 10, have been studied empirically using the plotwise information on geographical area and area under paddy in a few villages in West Bengal. The study relates to systematic sampling of 6 plots and the following three procedures xxx have been studied:

(i) linear systematic sampling and the usual unbiased estimator
(ii) linear systematic sampling and the biased ratio estimator
(iii) linear systematic sampling with probability proportional to the total size of the sample and the unbiased ratio estimator.

In procedure (iii) one plot is selected from the whole population with probability proportional to geographical area and the linear systematic sample containing this selected plot is selected. With this selection procedure the usual biased ratio estimator becomes unbiased. The results of this empirical study are given in Table 2. Though it has been shown that the unbiased ratio estimator is more efficient than the biased ratio estimator in case of large samples, from Table 2 it seems that this result is likely to be true for small samples also, though the gain does not seem to be substantial.

Table 2 - Showing the mean square errors of the estimators based on the three procedures considered in the empirical study.

village serial number	no. of plots	geographical area	area under paddy	bias of ratio estimator	mean square error		
					unbiased estimator	biased ratio estimator	unbiased ratio estimator
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			paddy variety (1)				
1	70	396.71	51.35		1306	953	815
2	122	586.12	48.10		1020	894	846
3	167	873.37	4.69		1136	153	131
4	103	652.56	92.25		1186	2687	2439
5	53	370.47	104.11		1111	1235	1059
6	22	152.90	34.30		81	191	163
			paddy variety (2)				
1	70	396.71	146.99		3214	3391	2917
2	122	586.12	417.95		5513	8536	8356
3	167	873.37	445.06		6741	8047	6413
4	103	652.56	129.03		2988	3497	2985
5	53	370.47	95.13		1229	1167	1043
6	22	152.90	40.92		332	99	75

12. ILLUSTRATIONS (III).

The efficiencies of different selection and estimation procedures considered here have been studied for sampling 2 units from the small population of 4 units given below. The results of this study are presented in Table 3.

Table 3 - Showing the efficiencies of different selection and estimation procedures.

sl. no.	selection procedure	estimator of $\bar{Y}$	bias	mean square error	efficiency
1	pps and srs of the remaining	$(\bar{y}/\bar{x})X$	-	0.34	88
2	equal probability without replacement	$N\bar{y}$	-	5.44	6
3	" " "	$(\bar{y}/\bar{x})X$	-0.15	0.39	87
4	" " "	Y'*	-	0.60	50
5	pps systematic	$\frac{1}{2}\sum_{i=1}^{2}\frac{y_i}{p_i}$	-	0.30	100

*$Y' = X + \frac{N-1}{n-1} n(y - r x)$

To study the efficiency of the ratio estimator empirically, the village-wise crop acreage data for three police stations (administrative units which are groups of villages) were used. The geographical area is taken as the supplementary variate and the characteristic under consideration as the area under a crop. Suppose the crop area is to be estimated by selecting a sample of villages. The sampling schemes considered are

(i) simple random sampling with replacement

(ii) probability proportional to area sampling with replacement

(iii) ratio estimation in case of srs with replacement

The variances of the estimators in the above three cases calculated by using the usual formulae. The variances are of the form $\frac{V}{n}$ where n is the sample size and the values of V are given below.

Table 4 - Showing the values of variance per unit for the three estimators under consideration.

crop	police station	Variance per unit (000) srs	pps	srs ratio estimator
Jute	Haringhata	1390	755	638
Hanskhali	Hanskhali	4640	3565	4235
	Santipur	12460	8620	8702
Aus	Haringhata	43270	17986	22639
	Hanskhali	304398	80102	44516
	Santipur	166255	43283	73994
Aman	Haringhata	90973	43125	49974

From the above table, it may be observed that the pps and ratio estimators are equally efficient and that they are more efficient than the srs estimator.

REFERENCES

1. GOODMAN, L. A. and Hartley, H.O. (1958): The precision of unbiased ratio type estimators, Jour. Amer. Stat. Assn., 53, 491-508.

2. HARTLEY, H.O. and ROSS, A. (1954) : Unbiased ratio estimators, Nature, 174, 270-271.

3. MURTHY, M. N. and NANJAMMA, N.S.(1959) : Almost unbiased ratio estimates based on interpenetrating sub-sample estimates, Sankhya, 21, 381-392.

4. MURTHY, M. N., NANJAMMA, N.S. and SETHI, V.K. (1959) : Some sampling systems providing unbiased ratio estimators, Sankhya, 21, 299-314.

5. SUKHATME, P.V. (1953) : Sampling Theory of Surveys with Applications, Iowa College Press, Chapter IV, 138-182.

Chapter 5

UNBIASED PRODUCT ESTIMATORS

1. INTRODUCTION

In this chapter a technique is developed to estimate the bias of an ordinary estimator of product of two population parameters on the basis of independent interpenetrating sub-sample estimates. This estimator of bias is used to correct the product estimator for its bias, thereby obtaining an unbiased product estimator. The condition under which the unbiased product estimator is more efficient than the biased estimator has been considered. This technique has also been extended to the case of estimating the product of several parameters unbiasedly. Incidentally the concept of using the product method of estimation to improve upon conventional estimators is introduced.

Though the problem of obtaining unbiased ratio estimators has received considerable attention in recent years, the question of getting unbiased estimator of product of population parameters or that of making product of unbiased estimators unbiased for the population parameter has received much less attention. There are a number of situations where the latter problem becomes important. For instance, in case of crop surveys, the estimate of production

is obtained as a product of the crop acreage estimator based on a suitable probability sample and the yield rate estimated from a sub-sample of this sample. Another situation where product of estimators is used is the construction of cost of living index and such other indices. This problem occurs frequently in case of multi-phase sampling where a ratio estimator or a chain of ratio estimators is usually used for estimating the population parameter.

In case of ratio method of estimation, Murthy and Najamma (1959) developed a technique of obtaining an almost unbiased ratio estimator based on interpenetrating sub-sample estimates. This technique has been generalized in chapter 2 with a view to obtain (almost) unbiased estimators for non-linear parametric functions. In this chapter this technique is applied to estimate the bias in a product estimator and this estimator of bias has been used to obtain an unbiased product estimator. It is interesting to note that almost all the results derived in the case of ratio estimators hold in this case also. First two types of product estimators are compared from the points of view of bias and mean square error. These two types of estimators are used in estimating the bias of the product estimators. The results obtained in the case of a product of two estimators have been generalized to the case of product of several estimators by considering a series of product estimators based on independent interpenetrating sub-sample estimates. It may

be noted that the treatment given here is quite general and applies to any probability design where the sample is drawn in the form of independent interpenetrating sub-samples.

As mentioned earlier, the concept of using product method of estimation to improve upon conventional estimators is introduced in section 3. This method consists in dividing the product of unbiased estimators of the parameter under consideration and of the parameter corresponding to an auxiliary variate by the population value of the auxiliary variate. The condition under which this product estimater is more efficient than the conventional unbiased estimator is also considered.

It may be mentioned that though all the results devised here can be obtained as particular cases of the results derived in chapter 2, a slightly different approach which is more direct has been adopted in this chapter.

2. BIAS AND MEAN SQUARE ERROR OF PRODUCT ESTIMATOR

Let t_1 and t_2 be unbiased estimators of the population parameters T_1 and T_2 respectively based on any probability sample. The product t_1t_2 $(=P)$ is usually considered as an estimator of the product T_1T_2 $(=T)$. This estimator is consistent, but biased. Writing $t_1 = T_1(1 + e_1)$ and $t_2 = T_2(1 + e_2)$, we get the bias and the mean square error of the product estimator P as

$$B(P) = E(t_1t_2 - T)$$

$$= T\, E\,[\,(1 + e_1)\,(1 + e_2) - 1]$$

$$= T\, v_{11} = T\, \rho\, C_1 C_2 \qquad (2.1)$$

and

$$M(P) = E(t_1 t_2 - T)^2$$

$$= T^2\, E(e_1 + e_2 + e_1 e_2)^2$$

$$= T^2\,(v_{20} + 2v_{11} + v_{02} + 2v_{12} + 2v_{21} + v_{22}) \qquad \ldots \qquad (2.2)$$

where

$$v_{ij} = \frac{E(t_1 - T_1)^i\,(t_2 - T_2)^j}{T_1^i\, T_2^j} \qquad (2.3)$$

C_1 and C_2 are coefficients of variation of the estimators t_1 and t_2 and ρ is the correlation coefficient between t_1 and t_2. In case of sampling schemes such as simple random sampling and varying probability sampling with replacement, the bias and mean square error of the product estimator P become

$$B(P) = T\, \frac{C_1'\, C_2'}{n} \qquad (2.4)$$

$$M(P) = T^2\,\Big[\frac{1}{n}\,(v'_{20} + 2v'_{11} + v'_{02}) + \frac{2}{n^2}\,(v'_{12} + v'_{21}) +$$

$$\frac{1}{n^3}\, v'_{22} + (n-1)v'_{20}\, v'_{02}\Big] \qquad (2.5)$$

where v'_{ij} stands for v_{ij} given in (2.3) for one sample unit and n is the sample size. This shows that if the sample size is large,

then the contribution of bias to mean square error wouod be negligible. If n is large the terms involving n^{-2} and n^{-3} in the mean square may be neglected and (2.5) reduces to

$$M(P) = \frac{T^2}{n}(v'_{20} + 2v'_{11} + v'_{02}) \qquad (2.6)$$

If t_1 and t_2 are uncorrelaod, then the bias is zero and the product estimator P is unbiased for T. Under the assumption that t_1 and t_2 are bivariate normally distributed, the mean square error of P becomes

$$M(P) = T^2 [\, C_1^2 + 2\, C_1C_2 + C_2^2 + (1 + 2/\ ) C_1^2 C_2^2 \,] \qquad (2.7)$$

3. PRODUCT METHOD OF ESTIMATION

It is interesting to note that the product method of estimation may be used to improve the estimator by using a suitable supplementary variate just as in the case of ratio method of estimation. SupposeT_2 is the parameter value corresponding to the supplementary variate and is known. Then the product estimator of T_1 which uses the supplementary information is given by

$$T_1 = \frac{t_1 t_2}{T_2} \qquad (3.1)$$

The bias of the estimator is given by

$$B(T_1) = T_1 v_{11} \qquad (3.2)$$

and the mean square error to the second degree approximation is

$$M(T_1) = T_1^2(v_{20} + 2v_{11} + v_{02}) \qquad (3.4)$$

For large sample the bias in this estimator is likely to be negli-

gible and this estimator will be more efficient than the estimator t_1 if

$$M(T_1) < T_1^2 v_{20}$$

that is, if

$$< -\tfrac{1}{2} \frac{C_2}{C_1} \qquad (3.4)$$

This shows that the product estimator is more efficient than the ordinary estimator if the estimators t_1 and t_2 are negatively correlated and if the correlation coefficient between these estimators is less than the expression given in (3.4). This result is of interest because it shows that the product estimator is likely to be more efficient wherever a ratio estimator turns out to be less efficient than the ordinary estimator. In fact for a given supplementary variate, one can decide whether to use ratio estimator, product estimator or ordinary estimator depending on the value of the correlation coefficient between the estimator of the variate under consideration and that of the supplementary variate. That is, the estimators $\frac{t_1 t_2}{T_2}$, t_1 and $\frac{t_1}{t_2} T_2$ are to be chosen according as the correlation coefficient between t_1 and t_2 is less than $-\frac{1}{2} \frac{C_1}{C_2}$, lies between $-\frac{1}{2} \frac{C_1}{C_2}$ and $+\frac{1}{2} \frac{C_1}{C_2}$. This will lead to better utilization of the available supplementary information in improving upon the ordinary estimator.

4. COMPARISON OF TWO PRODUCT ESTIMATORS.

Suppose in a survey the sample is drawn in the form of n independent interpenetrating sub-samples. Let (t_{1i}, t_{2i}) be unbiased estimators of T_1 and T_2 based on the ith sub-sample (i=1,2,..., n). The following two estimators can be considered to estimate the product T_1T_2 (=T).

$$P_1 = t_1t_2 = \left(\frac{1}{n}\sum_i^n t_{1i}\right)\left(\frac{1}{n}\sum_i^n t_{2i}\right) \tag{4.1}$$

$$P_n = \frac{1}{n}\sum_i^n t_{1i}\, t_{2i} \tag{4.2}$$

Applying result (2.1) to P_1, we get

$$B(P_1) = B_1 = \bar{Tv}_{11}$$

where $v_{11} = E(t_1 - T_1)(t_2 - T_2) = \frac{1}{n^2}\sum_i^n E(t_{1i} - T_1)(t_{2i} - T_2)$

since t_{1i} and t_{2j} $(j \neq i)$ are uncorrelated. Hence we have

$$B(P_1) = \frac{1}{n}\sum_i^n B(t_{1i}\, t_{2i}) \tag{4.3}$$

The bias of P_n is given by

$$B(P_n) = B_n = \frac{1}{n}\sum_i^n B(t_{1i}\, t_{2i}) \tag{4.4}$$

Comparing (4.3) and (4.4), we see that the bias of P_n is n times the bias of P_1, that is,

$$B(P_n) = nB(P_1) \tag{4.5}$$

If it is assumed that the sub-sample sizes are the same (as is generally the case), we get

$$B(t_{1i}\ t_{2i}) = B$$

$$M(t_{1i}\ t_{2i}) = M$$

$$v_{rs}\ (t_{1i}\ t_{2i}) = v_{rs} \text{ for all } i,$$

$$v_{rs} = \frac{v_{rs}}{n^{r+s}}\ ,\ r = 1, 2,\ \ s = 0, 1, 2,\ r+s = 2,3$$

and
$$v_{22} = \frac{1}{n^3}\left[v_{22} + (n-1)\ v_{02}\ v_{20} + 2\ v_{11}^2\right] \quad .. \quad (4.6)$$

By applying the result (2.2) to P_1 and making use of the results in (4.6) we get

$$M(P_1) = T^2\left[\frac{1}{n}(v_{20} + 2v_{11} + v_{02}) + \frac{2}{n^2}(v_{21} + v_{12})\right.$$

$$\left. + \frac{1}{n^3}\ v_{22} + (n-1)(v_{20}\ v_{02} + 2v_{11}^2)\ \right] \qquad (4.7)$$

$$= \frac{M}{n} - \frac{n-2}{n^2}\,A$$

where

$$A = T^2\left[2(v_{12} + v_{12}\ ) + \frac{n+1}{n}\,v_{22} - \frac{1}{n}(v_{20}\ v_{02} + 2v_{11}^2)\right] \qquad (4.8)$$

The mean square error of P_n is given by

$$M(P_n) = M_n = E(P_n - P)^2 = \left[\frac{1}{n}\,(t_{1i}\ t_{2i} - T)\right]^2$$

$$= \frac{M}{n} + \frac{n-1}{n}\,B^2 \qquad (4.9)$$

From (4.7) and (4.9), we have

$$M_n = M_1 + \frac{n-1}{n^2}\,A + \frac{n-1}{n}\,B^2 \qquad (4.10)$$

Comparison of M_n and M_1 in general difficult. Assuming that t_{1i} and t_{2i} are bivariate normally distributed, we get

$$A = T^2(1+2\ ) C_1^2 C_2^2 > 0 \qquad (4.11)$$

This shows that the mean square error of P_n is greater than that of P_1. Since $M_1 = V_1 + B_1^2$ and $M_n = V_n + B_n^2$, we get

$$V_n = V_1 + \frac{n-1}{n^2}(A - B^2) \qquad (4.12)$$

Under the assumption of bivariate normal distribution of t_{1i} and t_{2i},

$$A - B^2 = T^2(1 + \quad) C_1^2 C_2^2 > 0 \qquad (4.13)$$

which shows that P_1 is more efficient than P_n even from the point of view of variance. Thus we see that P_1 is more efficient than P_n from the points of view of bias, mean square error and variance. Since A and $(A-B^2)$ are independent of n, the difference between the efficiencies of P_1 and P_n decreases with increase in sample size.

5. ESTIMATION OF BIAS OF PRODUCT ESTIMATOR.

An unbiased estimator of the bias of the product estimators P_1 and P_n considered in section 4 can be obtained as given below.

$$E(P_1) = T + B_1 \qquad (5.1)$$

$$E(P_n) = T + B_n \qquad (5.2)$$

Subtracting (5.1) from (5.2), we get

$E(P_n - P_1) = B_n - B_1 = (n-1)B_1$, since $B_n = nB_1$. Hence an unbiased estimator of the bias of P_1 is given by

$$B_1 = \frac{P_n - P_1}{n-1} \tag{5.3}$$

and that of P_n will be nB_1. The variance of this estimator of bias is given by

$$V(B_1) = \frac{1}{(n-1)^2}\left[V_n + V_1 - 2\rho' \sqrt{V_1 V_n}\right]$$

where ρ' is the correlation coefficient between the estimators P_1 and P_n. Using (4.12), $V(B_1)$ may be written as

$$V(B_1) = \frac{n}{(n-1)^2}\left(a^2 - 2\rho' a + 1\right) \tag{5.4}$$

where

$$a^2 = 1 + \frac{n-1}{n^2}\,\frac{B^2 - A}{V_n} < 1.$$

6. UNBIASED PRODUCT ESTIMATOR

The unbiased estimator of bias of P_1 obtained in (5.3) may be used to correct the estimator P_1 for its bias, thereby obtaining an unbiased product estimator P_c given by

$$P_c = \frac{nP_1 - P_n}{n-1} \tag{6.1}$$

The variance of the corrected estimator is

$$V(P_c) = \frac{V_n}{(n-1)^2}\left(n^2a^2 - 2n\rho' a + 1\right) \tag{6.2}$$

where a^2 is as defined in (5.4). The gain in precision in using

P_c instead of P_1 is given by

$$G(P_c) = \frac{M_1 - V(P_c)}{M_1} = 1 - \frac{n^2a^2 - 2\rho a + 1}{(n-1)^2(a^2 + s^2)} \qquad (6.3)$$

where $s^2 = \frac{B^2}{n^2 V_n} = \frac{B^2}{nV}$, B and V being the bias and variance of the product estimator based on one sub-sample. A sufficient condition for $G(P_c) > 0$ is

$$(n-1)^2a^2 - (n^2a^2 - 2n\rho a + 1) > 0$$

that is, if

$$(2n-1)a^2 - 2n\rho a + 1 < 0 \qquad (6.4)$$

which will be true if a lies between the roots of the equation,

$$(2n-1)a^2 - 2n\rho a + 1 = 0.$$

The table showing for given values of ρ and a, the minimum value of n required to make the gain positive, the optimum n and the maximum gain, given by Murthy and Nanjamma (1959) is reproduced below for ready reference.

Table 1. Minimum and maximum values of $G(P_c)$ with the corresponding values of n for different ρ and a where $\rho > a$.

sr.no.	a	ρ	minimum		maximum	
			n	$G(P_c)$	n	$G(P_c)$
(0)	(1)	(2)	(3)	(4)	(5)	(6)
1	0.6	0.7	6	0.0089	10	0.0192
2		0.8	3	0.0556	4	0.0988
3		0.9	2	0.0889	3	0.3056
4	0.7	0.8	4	0.0113	7	0.0266
5		0.9	2	0.1020	3	0.1684
6	0.8	0.9	3	0.0469	4	0.0486

It may be noted that this table has been worked out neglecting s^2 in (6.3).

7. ESTIMATION OF PRODUCT OF SEVERAL PARAMETERS.

Suppose t_i is an unbiased estimator of the parameter T_i $(i = 1, 2, \ldots, k)$. An estimator of $\prod_{i=1}^{k} T_i$ is given by

$$P = \prod_{i}^{k} t_i \tag{7.1}$$

This estimator is biased and the bias is given by

$$B(P) = E\left(\prod_{i}^{k} t_i - \prod_{i}^{k} T_i\right).$$

Writing $t_i = T_i(1 + e_i)$, we get

$$B(P) = T\left[\prod_{i}^{k} (1 + e_i) - 1\right]$$

$$= T\left(\sum e_i + \sum e_i e_j + \ldots\right)$$

$$\text{Hence } B(P) = T \sum_{r=2}^{k} \sum_{r} v_{i_1 i_2 \ldots i_r} \tag{7.2}$$

where $\sum_{r}$ stands for summation over combinations of r estimators and

$$v_{i_1 i_2 \ldots i_r} = E(t_{i_1} - T_{i_1})(t_{i_2} - T_{i_2}) \ldots (t_{i_r} - T_{i_r}).$$

Suppose t_{ij} is an unbiased estimator of T_i based on the jth independent interpenetrating sub-sample $(j=1,2,\ldots, n)$. Let the number of subsamples be a multiple of 2, 3, ..., k-1*. Using the different combinations of the subsample estimates, we can construct, the following product estimators.

$$P_m = \frac{1}{m} \sum_{r=1}^{m} \left[\prod_{i=1}^{k}\right. \tag{7.3}$$

multiple in

for m = 1, 2, ... k-1, n. It may be noted that there are a number of ways of partitioning n sub-samples into m partitions of n/m sub-samples each and that the mean of the product estimators based on them may be taken as P_m.

The bias of P_m is given by

$$B(P_m) = B_m = T \sum_{r=2}^{k} \frac{A_r}{(\frac{m}{n})^{r-1}}, \quad m=1,2,\ldots, k-1, n \qquad (7.4)$$

$$= T \sum_{r=1}^{k-1} m^r y_{r+1}, \; y_{r+1} = \frac{A_{r+1}}{n^r}$$

$$R(P_m) = T + B_m, \; m = 1, 2, \ldots k-1,n$$

$$E(P_m - P_1) = B_m - B_1 = \sum_{r=1}^{k-1} (m^r-1)\, y_{r+1}, \; m = 2, 3, \ldots k-1,n \qquad (7.5)$$

Writing $D_m = P_m - P_1$, we get from (7.5)

$$E(D) = (y)\,(\;\;)$$

where

$$B = (D_2, D_3, \ldots D_{k-1}, D_n)$$

$$y = (y_2, y_3, \ldots, y_{k-1})$$

$$\Lambda = \begin{pmatrix} 2-1 & 3-1 & \ldots & (k-1)-1 & n-1 \\ 2^2-1 & 3^2-1 & & (k-1)^2-1 & n^2-1 \\ \vdots & \vdots & & \vdots & \vdots \\ 2^{k-1}-1 & 3^{k-1}-1 & & & \end{pmatrix}$$

$$y = E(D) \tag{7.7}$$

From (7.4) we have

$$(B) = (y)\,(\quad) \tag{7.8}$$

where

$$(B) = (B_2, B_3, \ldots, B_{k-1}, B_n)$$

and is a (k-1) x (k-1) matrix in which all the elements are unity

$$(B) = E(D) + E(D)$$

where is given by

$$\begin{matrix} s_1 & s_1 \cdots & s_1 \\ s_2 & s_2 \cdots & s_2 \\ s_{k-1} & s_{k-1} \cdots & s_{k-1} \end{matrix}$$

where s_m is the sum of the elements in the mth row of . An unbiased estimator of (B) is given by

$$(\hat{B}) = (D) + (D)\,(\quad) = (D) + (D)(S)(1) \tag{7.9}$$

where $(S)' = (s_1, s_2, \ldots, s_{k-1})$ and $(1) = (1, 1, \ldots, 1)$. Hence

$$\hat{B}_m = \sum_j D_j S_{j-1} + D_m,\ j = 2, 3, \ldots, k-1, n,\ s_{n-1} \quad \ldots \tag{7.10}$$

for m = 2, 3, ..., k-1, n. Hence the corrected estimator is given by

$$P_c = P_m - \hat{B}_m \tag{7.11}$$

Particular cases:

(i) k = 2 $\quad P_1 = \bar{t}_1 \bar{t}_2$, $P_n = \frac{1}{n}\sum_i^n t_{1i}\, t_{2i}$

$$B_m = my_2 = \frac{m}{n} A_2, \quad m = 1, n$$

$$= (n-1) \qquad = \frac{1}{(n-1)} \quad (S) = \frac{1}{(n-1)}$$

Hence

$$B_n = (P_n - P_1) + (P_n - P_1) \frac{1}{(n-1)} = \frac{n}{(n-1)} (P_n - P_1)$$

and since $B_n - B_1 = E(P_n - P_1)$, $B_1 = \dfrac{P_n - P_1}{(n-1)}$

(ii) $k = 3$. In this case we consider the following three estimators.

$$P_1 = \bar{t}_1 \bar{t}_2 \bar{t}_3$$

$$P_2 = \frac{1}{k} \sum_{i=1}^{2} \pi_{i=1}^{3} \left(\frac{2}{n} \sum_{j=(r-1)\frac{n}{2}+1}^{r\frac{n}{2}} t_{ij} \right)$$

$$P_n = \frac{1}{n} \sum_{i=1}^{n} t_{1i} t_{2i} t_{3i}$$

Substituting the relevant values in appropriate expressions in section 7, we get the following estimators of the bias of P_1, P_2 and P_n

$$B_1 = -\frac{(n+1)}{(n-1)} P_1 + \frac{n}{(n-2)} P_2 - \frac{2}{(n-1)(n-2)} P_n$$

$$B_2 = -\frac{2n}{(n-1)} P_1 + 2\frac{(n-1)}{(n-2)} P_2 - \frac{2}{(n-1)(n-2)} P_n$$

$$B_n = -\frac{2n}{(n-1)} P_1 + \frac{n}{(n-2)} P_2 + \frac{n(n-3)}{(n-1)(n-2)} P_n$$

Hence an unbiased estimator of T is given by

$$P_c = \frac{2n}{(n-1)} P_1 - \frac{n}{(n-2)} P_2 + \frac{2}{(n-1)(n-2)} P_n.$$

8. RATIO CUM PRODUCT ESTIMATORS

In the previous sections, the problem of making a product of two or more unbiased estimators unbiased for the population parameter has been discussed. In this section, the question of making the product of an ordinary ratio estimator and an unbiased estimator unbiased for the parameter is considered. As has been pointed out earlier, this type of problem arises in estimation of crop production. In this case the production estimator is a ratio cum product estimator since it is obtained as a product of the crop acreage estimator based on a probability sample and the yield rate estimator based on a sub-sample of the original sample. The yield-rate estimator, being a ratio is based. This situation also occurs in multi-phase sampling where a ratio estimator is used.

Suppose t_1, t_2 and t_3 are unbiased estimators of the parameters T_1, T_2 and T_3 respectively and it is intended to estimate $\frac{T_1}{T_2} T_3 (=T)$. An estimator of this is given by

$$P' = \frac{t_1}{t_2} t_3 \qquad (8.1)$$

This estimator may also arise in case of two-phase sampling where t_1 is unbiased for T_1 and t_2 and t_3 are unbiased estimators of the supplementary variate parameter T_2 based on the second and first phase samples respectively. In this case T_2 and T_3 would be the same.

Writing $t_i = T_i(1 + e_i)$, $i = 1, 2, 3$, we get the bi s of P' as

$$B(P') = E\left(\frac{t_1}{t_2} t_3 - T\right)$$

$$= TE\left[\frac{(1+e_1)(1+e_3)}{(1+e_2)} - 1\right]$$

$$= T E\left[(e_1 + e_3 - e_2) + (e_2^2 + e_1 e_3 - e_1 e_2 - e_2 e_3) + \text{terms of degree greater than } 2\right],$$

assuming that $|e_2| < 1$. If the sample size is large, terms of degree greater than 2 in e_1, e_2 and e_3 in the above expression may be neglected and we get

$$B(P') = T(\nu_{200} + \nu_{101} - \nu_{110} - \nu_{011}) \tag{8.2}$$

where

$$\nu_{ijk} = E\,(t_1 - T_1)\,(t_2 - T_2)^j (t_3 - T_3)^k$$

Let the sample be drawn in the form of n independent interpenettrating sub-samples and let t_{ij} $(i = 1, 2, 3, j = 1, 2, \ldots, n)$ be the estimator of T_i based on the jth sub-sample. Using these sub-sample estimates we can construct the following two estimators.

$$P'_1 = \frac{t_1}{t_2} t_3 \tag{8.3}$$

$$P'_n = \frac{1}{n} \sum_{j}^{n} \frac{t_{1i}}{t_{2i}} t_{3i} \tag{8.4}$$

From (8.2) the bias of P'_1 correct upto the second degree of approximation is given by

$$B(P'_1) = T(\nu_{200} + \nu_{101} - \nu_{110} - \nu_{001})$$

where

$$\nu_{ijk} = E\,(t_1 - T_1)^i (t_2 - T_2)^j (t_3 - T_3)^k$$

Hence $$B(P'_1) = \frac{1}{n^2} \sum_i^n T(v_{200} + v_{101} - v_{110} - v_{011})$$

$$= \frac{1}{n^2} \sum_i^n B\left(\frac{t_{1i}}{t_{2i}} t_{3i}\right) \qquad (8.5)$$

The bias of P'_n is given by

$$B(P'_n) = \frac{1}{n} \sum_i^n B\left(\frac{t_{1i}}{t_{2i}} t_{3i}\right) \qquad (8.6)$$

Comparing the biases of P'_1 and P'_n given in (8.5) and (8.6) we see that

$$B(P'_n) = n\, B(P'_1) \qquad (8.7)$$

to the second degree of approximation. Hence and unbiased estimator of $B(P'_1)$ to the second degree of approximation is given by

$$B(P'_1) = \frac{P'_n - P'_1}{(n-1)} \qquad (8.8)$$

Using this estimator of bias of P'_1 an estimator P'_c can be obtained which will be unbiased to the second degree approximation. The estimator P'_c may

$$P'_c = \frac{nP'_1 - P'_n}{(n-1)} \qquad (8.9)$$

may be considered almost unbiased for the parameter under consideration.

9. AN ILLUSTRATION.

The technique of making a product estimator unbiased using subsample estimates has been applied to the estimates of crop production

and crop acreage given in Report Number 38 of the National Sample Survey. The sampling design in this survey consisted of a stratified two-stage design for the land utilization survey where villages were the first stage units and clusters of ten plots were the second stage units. The crop-cutting survey was confined to only one-third of the total sample. The crop production estimate is obtained as a product of the yield rate based on the one-third sample where crop cutting experiments were conducted and the crop acreage estimate based on the whole sample. In the report estimates of crop acreage and production have been given for two independent interpenetrating sub-samples. Suppose a_1, a_2 and y_1, y_2 are the estimates of crop acreage and yield rate based on sub-samples 1 and 2 respectively, then the two estimators given in section 2 are given by

$$P_1 = \frac{1}{4}(a_1 + a_2)(y_1 + y_2) \qquad (9.1)$$

$$P_2 = \frac{1}{2}(a_1 y_2 + a_2 y_2) \qquad (9.2)$$

The estimator corrected for its bias is given by

$$P_c = (2P_1 - P_2) \qquad (9.3)$$

The combined estimator given in the report (P, say) has been obtained from the sub-sample estimates by weighting them by the number of surveyed villages at stratum level.

The values of the estimators P_1, P_2, P_c and P are presented in Table 2 for different cereal crops at zonal level. From this

table it is clear that the bias in the estimator P is negligible except for certain minor crops.

Table 2 - Showing the values of the crop production estimators P_1, P_2, P_c and P. ('000 tons).

zone	no. of sample villages	P_1	P_2	P_c	P	P_1	P_2	P_c	P
(0)	(1)	(2)	(3)	(4)	(5)	(2)	(3)	(4)	(5)
		rice				jowar			
North India	445	841	843	839	841	305	294	316	297
Central India	692	3377	3378	3376	3310	2598	2622	2574	2383
East India	721	12264	12229	12299	12360	52	44	60	47
South India	529	7553	7555	7551	7639	2907	2910	2904	2634
West India	676	3551	3548	3554	3559	8328	8320	8336	8515
All India	3063	27578	27552	27604	27709	14189	14188	14190	13876

* Crop cutting in all seasons and land utilization in autumn season only in one-third of the villages.

		bajra				ragi			
North India	445	1257	1278	1236	1279	-	-	-	-
Central India	692	1276	1270	1282	1288	1	1	1	1
East India	721	33	25	41	25	328	340	316	381
South India	529	759	762	756	753	950	938	962	936
West India	676	1978	1970	1986	1978	829	836	822	865
All India	3063	5301	5305	5297	5323	2117	2114	2120	2183

Table 2 (Continued)

(0)	(1)	(2)	(3)	(4)	(5)	(2)	(3)	(4)	(5)
		maize				hwheat			
North India	445	2709	2708	2710	2703	4328	4320	4336	4458
Central India	692	1218	1202	1234	1224	5201	5194	5208	5239
East India	721	581	598	564	554	378	378	378	376
South India	529	156	156	156	153	10	12	8	12
West India	676	630	631	629	632	969	971	967	968
All India	3063	5314	5296	5332	5266	10879	10876	10882	11053
		barley				all cereals			
North India	445	649	652	646	637	10103	10096	10110	10215
Central India	692	1636	1660	1612	1661	15328	15327	15329	15106
East India	721	332	337	327	345	13984	13950	14018	14088
South India	529	-	-	-	-	12337	12334	12340	12127
West India	676	9	9	9	11	16286	16284	16288	16528
All India	3063	2654	2659	2649	2654	68005	67991	68019	68064

Source of data on which this table is based : National Sample Survey (1960) 'Some Results of the Land Utilization and Crop Cutting Experiments' , Report Number 38, issued by the Cabinet Secretariat, Government of India).

10. EMPIRICAL STUDY.

The plot-wise data on geographical area and area under paddy in a few villages in West Bengal are utilized to study empirically the efficiencies of ratio and product estimators. The study relates to systematic sampling of 6 plots and the following three estimators have been studied:-

(i) simple unbiased estimator

(ii) ratio estimator with geographical area as supplementary information

(iii) product estimator with geographical area x as supplementary information.

The results of this study are given in Table 3.

From Table 3, it can be seen that there are cases where the product estimator is more efficient than the ratio estimator. This example is given more by way of illustration than as a suggestion for use of product estimator in such situation. It may be noted that in case of crop survey if net cultivated area for a previous year is used as supplementary information, then a ratio estimators for major crops and product estimators for minor crops are likely to be more efficient than simple unbiased estimators.

REFERENCE

MURTHY, M.N. and NANJAMMA, N.S. (1959) : Almost unbiased ratio estimates based on interpenetrating sub-sample estimates, Sankhya, 21, 381-392.

Table 3 - Showing the mean square errors of the three estimators mentioned in the earlier page.

village sr.no.	no. of plots	geographical area	area under paddy	bias		mean square error		
				ratio	product	unbiased	ratio	product
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				Paddy variety (1)				
1	100	657.67	230.34	7.87	0.34	2859	4286	5687
2	132	804.58	238.57	-1.94	6.30	6590	4362	10782
3	232	1495.26	163.87	30.66	-8.91	11681	19773	11541
4	172	1116.85	260.94	10.12	-0.96	11380	14117	13122
5	79	389.12	78.44	0.05	2.20	1821	1670	2302
6	48	274.26	50.94	-0.15	1.49	536	488	852
7	91	570.22	100.30	1.54	1.36	2031	1942	3141
8	94	690.58	83.49	7.11	-3.36	2656	4256	2059
9	162	864.40	146.31	5.22	-1.18	3526	4071	4231
10	178	923.62	339.97	0.57	-1.55	8669	7863	12429
11	172	1195.00	227.43	9.78	-5.45	10855	14261	9713
12	233	1158.01	613.25	17.17	-4.49	42698	49170	48837
13	114	680.80	288.56	2.56	2.75	15803	14636	20250
14	118	737.73	286.03	-5.47	12.49	14727	9458	23629
15	53	319.54	126.87	-0.65	1.00	2791	2718	2943
16	98	910.31	354.05	14.35	11.03	24496	19839	49586
17	79	543.21	214.27	-0.90	5.61	8126	6632	12395
18	57	332.79	105.41	0.62	2.85	1290	989	2596
19	57	461.67	222.21	-0.19	1.87	3544	2846	3928
20	232	1322.83	704.28	1.72	24.73	59632	38740	152154
21	115	677.62	395.52	2.45	8.11	11784	8886	28383
				Paddy variety (2)				
1	100	657.67	60.37	0.86	0.87	1108	1118	1282
4	172	1116.85	53.58	1.15	2.11	1361	1315	2861
5	79	389.12	122.41	-1.20	4.93	2472	1594	4336
6	48	274.26	14.91	-1.55	2.22	225	125	429
7	91	570.22	35.33	4.98	-3.91	2845	3915	2172
9	162	864.40	216.69	3.92	3.30	3910	10676	8516
10	178	923.62	135.45	-0.69	1.24	5142	3774	6567
11	172	1195.00	383.87	-0.87	4.18	21167	16570	32173
12	233	1158.01	3.16	-0.38	0.54	90	60	159
13	114	680.80	1.33	-0.17	0.21	33	15	31
14	118	737.73	181.01	5.05	0.70	5799	7053	7164
16	98	910.31	9.02	0.36	-0.08	82	100	74
17	79	543.21	13.55	0.95	-0.63	228	282	207

Chapter 6

VARIANCE AND CONFIDENCE INTERVAL ESTIMATION

1. INTRODUCTION

Some of the results of the investigations on (i) methods of estimation of variance of estimate and (ii) methods of setting up confidence interval for a population parameter are given in this chapter. As there is an abundance of literature on these two topics, it may not be out of place here to give a brief review of the work that has already been done. This review is by no means exhaustive.

The aim here is to study the efficiencies of methods of estimation of variance and of setting up confidence intefvals which are operationally convenient. Invariably such methods are less efficient than the conventional methods involving much calculation at the stage of analysis. Sometimes it may be possible to strike a balance between the efficiency aimed at and the labour involved.

In this chapter a procedure of determining the sample size is given together with some specimen tables giving the sample sizes for different situations. The idea here is to fix the sample size in such a way that the probability of the length of the confidence interval for the parameter associated with a specified confidence coefficient being less than a given value is a pre-specified quantity. The tables giving values of the sample size for different values of the constants involved are under preparation. As mentioned earlier some specimen tables are given here.

In the case of simple random sampling from a normal population the variance of the estimate of μ, the mean in the population involves the parameter σ, the population standard deviation. Let a simple random sample of size N be drawn from a normal population with mean μ and standard deviation σ. Let the observations be $X_1, X_2 \ldots\ldots X_N$. The minimum variance estimate among the class of unbiased estimates of μ is $\bar{X}$, the sample mean and its standard error is $\sigma/\sqrt{N}$. A list of estimates of σ available in statistical literature is given below.

$$s_1 = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}} \qquad (2.1)$$

$$s_2 = \frac{1}{c_2}\sqrt{\frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{m}} \qquad (2.2)$$

$$s_3 = \frac{1}{c_2}\sqrt{\frac{\sum_{i=1}^{m}(\bar{x}_i - \bar{X})^2}{m}} \qquad (2.3)$$

$$s_4 = \frac{\bar{w}}{d_2} \qquad (2.4)$$

$$s_5 = \frac{\lambda_2 - \lambda_1}{u_2 - u_1} \qquad (2.5)$$

where $\bar{x}$ is the mean of a sub-sample of size m,

$\bar{x}_i$ is the mean of the i th random group with m observations such that $nm = N$, $(i = 1, 2, \ldots. n)$;

$\bar{w}$ is the mean of the ranges in n sub-groups of n elements in each;

λ_1 and λ_2 are two numbers to be properly chosen

u_1 and u_2 are given by

where p and q are proportions of the observations less than and between and respectively,

$$c_2 = \frac{2}{n} \cdot \frac{\left(\frac{n}{2}\right)}{\left(\frac{n-1}{2}\right)}$$

and $d_2 =$

where and x_1 is the smallest observation in the sample.

As is to be expected of these estimates, s_1 is the most efficient and the most difficult to calculate. In sampling from Normal population, the estimates s_2 and s_3 have the same efficiency. Hansen, Hurwitz and Madow (1953) have compared the variances of the estimates of the type s_2^2 and s_3^2 and observe that s_2^2 is more or less precise than s_3^2 according as is less than or greater than 3 where . The expressions for the variance of s_2^2 and s_3^2 are

$$V(s_2^2) = \left(\quad - \frac{n-3}{n-1} \right) \frac{\quad}{n} \qquad (2.6)$$

$$V(s_3^2) = \left(\quad - \frac{n-3}{n-1} \right) \frac{\quad}{n} \qquad (2.7)$$

where . Hence it follows that $V(s_2^2)$ $V(s_3^2)$ according as 3.

The values of c_2 and d_2 are tabulated for different values of n and n in ASTM manual (1951). If the size of the sub-group is small (about 7 or 8 observations), then the loss of efficiency in using s_4 instead of s_1 as an estimate of is not large (Pearson and Haines

(1935)). Pearson (1932) has tabulated the mean, standard deviation and percentage limits (0.5%, 1%, 5% and 10%) of range in samples from a normal population for sample sizes 2(1) 30(5) 100. Cadwell (1954) has given an asymptotic expression for the probability integral of range of samples from a symmetrical unimodel population and has studied its accuracy for the case of normal parent population and for sample sizes 20 to 100. Stevens (1948) suggested the estimate s_5 and he has tabulated the efficiency of this estimate as compared to that of s_1 in large samples for different values of and , while sampling from a normal population with mean and standard deviation .

An empirical study was conducted to study the efficiency of s_5 as compared to that of s_1 for a sample of size 100 from a normal population. For this purpose the samples from the normal population with mean 0 and standard deviation 1 given by Mahalanobis and others (1934) have been used. There are 104 samples of size 100. For each of these the mean, standard deviation and frequency distribution have been given. The mean and variance of the sample standard deviations are 0.9887 and 0.0049 respectively. Taking λ_1 and λ_2 to be -0.5 and 0.5, for each sample s_5 was calculated. The mean and variance of s_5 turned out to be 1.0009 and 0.0193 respectively. Hence s_5 can be considered to be unbiased for this sample size and the efficiency of s_5 as compared to that of s_1 is 25% which agrees with the figure given by Stevens. The efficiency of s_5 can be increased by taking the values of and near about the mean on either side of it.

3. INTERPENETRATING SUB-SAMPLES

In a stratified sampling design where n independent and interpenetrating subsamples are taken from each stratum according to any sampling design, estimates of the variance of the estimate can be got by using (i) the subsample estimates of total or (ii) the subsample estimates of strata totals. It is of interest to get an expression for the loss of efficiency in using the former in preference to the latter.

Let there be k strata and n independent and interpenetrating subsamples in each stratum. For the sake of simplicity let the subsamples sizes within each stratum be the same. Suppose y_{ij} is an unbiased estimate of the j th stratum total Y_j from the i th subsample ($i = 1,2, \ldots n$; $j = 1,2, \ldots k$). The two estimates of the variance of the estimate y of the total Y are

$$(i)\quad V_1(y) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (y_{i.} - y)^2 \qquad (3.1)$$

and

$$(ii)\quad V_2(y) = \frac{1}{n(n-1)} \sum_{j=1}^{k} \sum_{i=1}^{n} (y_{ij} - Y_j)^2 \qquad (3.2)$$

where $y_{i.} = \sum_{j=1}^{k} y_{ij}$, $Y_{.j} = \frac{1}{n} \sum_{i=1}^{n} y_{ij}$ and $y = \frac{1}{n} \sum_{i=1}^{n} y_{i.}$.

It can be easily verified that the above two estimates of the variance are unbiased. The variances of the two estimates are given by

$$V\,V_1(y) = \frac{1}{n^3(n-1)} \sum_{j=1}^{k} (n-1)\,\mu_{4j} + (3-n)\,\mu_{2j}^2 + 4n \sum_{j=1}^{k} \sum_{i\,j} \qquad (3.3)$$

and $V\ V_2(\bar{Y}) = \frac{1}{n^3(n-1)} \sum_{j=1}^{k} (n-1)\ \mu_{4j} + (3-n)\ \mu_{2j}^2$ (3.4)

where μ_{2j} and μ_{4j} are the second and fourth moments of the estimate $\bar{y}_{ij}$. From the above expressions it follows that $V\ V_1(\bar{Y}) \geq V\ V_2(\bar{Y})$. The loss of efficiency in using $V_1(\bar{y})$ instead of $V_2(\bar{y})$ as an estimate of $V(\bar{y})$ is given by

$$L = \frac{V\ V_1(\bar{Y}) - V\ V_2(\bar{Y})}{V\ V_2(\bar{Y})} = \frac{n}{n-1} \cdot \frac{4 \sum_{j=1}^{k} \sum_{i<j} \mu_{2i}\ \mu_{2j}}{\sum_{j=1}^{k} \left(\beta_j - \frac{n-3}{n-1}\right) \mu_{2j}^2} \qquad (3.5)$$

where $\beta_j = \frac{\mu_{4j}}{\mu_{2j}^2}$. If the distribution of the estimates within each stratum can be assumed to be normal, then $\beta_j = 3$ for all j. Hence L becomes

$$L = \frac{2 \sum_{i<j} \mu_{2i}\ \mu_{2j}}{\sum_{j=1}^{k} \mu_{2j}^2} = \frac{\mu_2^2}{\sum_{j=1}^{k} \mu_{2j}^2} - 1 \qquad (3.6)$$

where $\mu_2 = \sum_{j=1}^{k} \mu_{2j}$. If the coefficient of variation of the estimate in each of the strata can be assumed to be equal, then L is given by

$$L = \frac{2 \sum_{i<j} Y_i^2\ Y_j^2}{\sum_{j=1}^{k} Y_j^4} . \qquad (3.7)$$

Instead, if it is assumed that the variance of the estimate in each stratum is the same, then L is equal to k-1. It may be noticed that the loss may be substantial if the number of strata is large.

4. CONFIDENCE INTERVAL ESTIMATION

If a sample of size N is drawn from a normal population with mean , and standard deviation , then the confidence interval for μ is given by

$$P\left[\bar{X} - t\frac{}{\sqrt{N}} \qquad \bar{X} + t\frac{}{\sqrt{N}}\right] = 1 - \qquad (4.1)$$

where 1- is the confidence coefficient and t is the % point of the distribution of $\frac{\bar{X} - }{}\sqrt{N}$. In practice one has to estimate from the sample itself by one of the procedures given in §4.

If s_1 is taken as an estimate of , then it is well-known that the statistic

$$t = \frac{(\bar{X} -)}{s_1}\sqrt{N-1}$$

is distributed as Student's t with N-1 degrees of freedom. Of course for large samples the above statistic is distributed normally with mean 0 and standard deviation 1. Similarly the statistic

$$t' = \frac{(\bar{X} -)}{s_2} \quad \text{and} \quad t'' = \frac{(\bar{X} -)}{s_3}$$

are also distributed as Student's t with (m-1) degrees of freedom, where m is the number of groups or subsample size and and are constants.

Daly (1946) has proved that $\bar{x}$ and w, the mean and range of sample of N independent observations on a normally distributed variate x are statistically independent. Lord (1947) has given the 5% and 1% points of the distribution of the statistic

$$u = \frac{(\bar{X} -)}{\bar{w}} d_2\sqrt{n\,m} \qquad (4.2)$$

where n is the sub-group size and m the number of sub-groups. Patnaik (1950) has obtained an approximation to the distribution of $\bar{w}$ and making use of this has derived the distribution of u. Jackson and

Ross (1955) have transformed the tables of Lord so as to provide the percentage points of the distribution of the statistic

$$G = \frac{(\bar{X} - \quad)}{\bar{w}} = \frac{u}{d_2\sqrt{nm}} \tag{4.3}$$

Hoether (1955) has considered the statistics

$$G_1 = \frac{(\bar{X} - \mu)}{\bar{w}} \quad \text{and} \quad G_2 = \frac{\bar{X}_1 - \bar{X}_2}{\bar{w}^1}, \tag{4.4}$$

where $\bar{w}^1$ is the mean of the ranges of all subgroups of both the samples, and has given the percentage points for G_1 and G_2 so that confidence intervals for and $(\quad_1 - \quad_2)$ can be set up in the form

$$P\ \bar{X} - \quad \bar{w} \qquad \bar{X} + \quad \bar{w} = 1 - \tag{4.5}$$

and $$P\ (\bar{X}_1 - \bar{X}_2) - \quad w' \tag{4.6}$$

where and are the . points of the distributions of G_1 and G_2. Further he has tabulated the values of a_N for different values of n the subgroup size and m the number of subgroup ($nm = N$) which when multiplied by the sum of the ranges in the subgroups provides us an unbiased estimate of .

Let $x_1\ x_2 \ldots\ldots x_n$ be independent observations on a variate x with some distribution function arranged in the increasing order of magnitude. Thompson (1936) has shown that

$$P(X_k \quad M \quad X_{n-k+1}) = 1 - 2\, I_{0.5}(n-k+1, k) \tag{4.7}$$

where M is the median in the population and Ix (p,q) is the incomplete Beta function which has been tabulated by Karl Pearson. If the distribution of x is symmetrical, then the above expression gives us the confidence region for the population mean. Nair (1940) has tabulated the values of k which give us confidence

intervals with confidence coefficient greater than or equal to 0.95 and 0.99 for values of n = 6(1) 81.

In what follows the efficiencies of the confidence intervals based on s_1 and s_2 will be compared. For the sake of convenience let us redefine s_1^2 and s_2^2 as

$$s_1^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} (X_i - \bar{X})^2 \qquad (4.8)$$

and $$s_2^2 = \frac{1}{N(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (4.9)$$

where x_1 x_2 ..... x_N are the N observations drawn from a normal population with mean and standard deviation and n is the subsample size. It is clear that $N(N-1)\frac{s_1^2}{2}$ and $N(n-1)\frac{s_2^2}{2}$ are distributed as with N-1 and n-1 degrees of freedom respectively. Hence it follows that the statistics

$$t_1 = \frac{(\bar{X} -)}{s_1} \quad \text{and} \quad t_2 = \frac{(\bar{X} -)}{s_2}$$

are distributed as student's t with N-1 and n-1 degrees of freedom respectively. If t and t_2 are the limits of the distribution of t_1 and t_2, then the lengths of the confidence intervals by the two methods will be

$$L_1 = 2t_1 \quad s_1 \qquad \text{and} \qquad L_2 = 2t_2 \quad s_2 \qquad (4.10)$$

A number of criteria can be suggested for comparing the coefficiencies of L_1 and L_2. L_1 and L_2 may be said to have approximately the same efficiency if $E(L_2)$ and $V(L_2)$ are nearly equal to $E(L_1)$ and $V(L_1)$ respectively. It is to be noted that $E(L_2)$ and $V(L_2)$ tend to $E(L_1)$ and $V(L_1)$ respectively as n tends to N. But the convergence after a certain stage becomes slow in the case of the expected value. The expected values are given by

$$E(L_1) = 2t_1 \cdot \frac{\sigma}{\sqrt{N}} \qquad \text{if N is large (} \quad 25) \qquad (4.11)$$

$$\text{and } E(L_2) = 2t_2 \, C_2' \frac{\sigma}{\sqrt{N}} \qquad \text{where } C_2' = \sqrt{\frac{n}{n-1}} \, C_2 . \qquad (4.12)$$

If N is fairly large (100) then t_1 = 1.96, for in that case t_1 is distributed normally with mean 0 and standard deviation unity. Table 1 gives the values of the ratio $E(L_2)/E(L_1)$ for different values of n, assuming N to be large.

Table 1 - Values of the ratio of the expected value of L_2 to that of L_1 for different values of n.

n	2	3	4	5	6	7	8	9	10	15	20	25
$\frac{E(L_2)}{E(L_1)}$	5.172	1.946	1.496	1.331	1.248	1.198	1.164	1.141	1.123	1.075	1.054	1.042

The confidence interval L_1 and L_2 may be said to have approximately the same efficiency if L_2^* is nearly equal to L_1^* where L_1^* and L_2^* are given by

$$P \quad L_1 \quad L_1^* \quad = 0.95 \qquad (4.13)$$

$$P \quad L_2 \quad L_2^* \quad = 0.95 . \qquad (4.14)$$

This criterion is defective in the sense that even if L_2^* is nearly equal to L_1^* at this level of confidence, this may not be true for some other level. A better approach may be to compare the distribution functions of L_1 and L_2 for different values of n. Here also it may be observed that the convergence of the distribution function of L_2 to that of L_1 is likely to become very a slow for values of n greater than a certain value. Table 2 gives the values of L_1^* and L_2^* for different values of N and n. Table 3 shows the distribution function of L_1 for N = 100 and that of L_2 for n = 10, 20 and 40.

Table 2 - Comparison of the values of L_1^* and L_2^* for different values of N and n at 95% confidence level.

N	L_1^*	values of L_2^*						
		n=2	n=3	n=4	n=5	n=6	n=8	n=40
96	0.223	2.559	0.750	0.514	*	0.380	0.332	*
120	0.198	2.273	0.680	0.469	0.380	0.349	0.306	0.283
200	0.150	1.761	*	0.363	0.302	*	0.237	0.219

Table 3 - Comparison of the distribution functions of L_1 and L_2 for N = 100, n = 4,5,10,20 and 40

λ	$p(L_1 < \lambda)$	$p(L_2 < \lambda)$				
		n = 4	n = 5	n = 10	n = 20	n = 40
0.22	0.0000	0.1294	0.1298	0.1053	0.0599	0.0200
0.24	0.0040	0.1644	0.1734	0.1701	0.1379	0.0832
0.26	0.1128	0.1998	0.2191	0.2541	0.2570	0.2304
0.28	0.5902	0.2352	0.2709	0.3541	0.4111	0.4551
0.30	0.9522	0.2760	0.3278	0.4591	0.5772	0.6930
0.32	0.9991	0.3218	0.3833	0.5621	0.7252	0.8655
0.34	1.0000	0.3663	0.4416	0.6641	0.8401	0.9547

As it is easier to compute s_2 then s_1, the object should be to find subsample size required to give us L_2 which does not differ much from L_1. In other words the subsample size should be so chosen that the variation of L_2 about L_1 is not much keeping an eye on the labour involved. L_2 may be said to be approximately as efficient as L_1, if

$$P\left(\frac{L_2}{L_1} - 1 \right) \qquad (4.15)$$

is fairly large where is a small quantity. To find this probability we require the distribution of (L_2/L_1).

Theorem: If $x_1\ x_2 \ldots\ldots x_N$ be N observations on a variate x which is normally distributed with mean and standard deviation , then the distribution of

$$Z = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{\sum_{i=1}^{N}(x_i - \bar{x}_N)^2} \qquad (4.16)$$

where $\bar{x}_n$ is the mean of the first n observations and $\bar{x}_N$ is the mean of all the N observations is that of a Beta variate with parameters $\frac{n-1}{2}$ and $\frac{N-n}{2}$

Proof:

$$\sum_{i=1}^{N}(x_i - \bar{x}_N)^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + \sum_{i=n+1}^{N}(x_i - \bar{x}_{N-n})^2 + (N-n)(\bar{x}_{N-n} - \bar{x}_N)^2$$

where $\bar{x}_{N-n} = \frac{1}{N-n}\sum_{i=n+1}^{N} x_i$

since $\bar{x}_N = \frac{n\bar{x}_n + (N-n)\bar{x}_{N-n}}{N}$

$\bar{x}_n - \bar{x}_N = \frac{N-n}{N}(\bar{x}_n - \bar{x}_{N-n})$ and $\bar{x}_{N-n} - \bar{x}_N = \frac{n}{N}(\bar{x}_{N-n} - \bar{x}_N)$

Hence $n(\bar{x}_n - \bar{x}_N)^2 + (N-n)(\bar{x}_{N-n} - \bar{x}_N)^2 = \frac{n(N-n)}{N}(\bar{x}_n - \bar{x}_{N-n})^2$

which is a with one degree of freedom, for

$$\sqrt{\frac{n(N-n)}{N}}\ (\bar{x}_n - \bar{x}_{N-n}) \quad \text{is } N(0,1)$$

Further $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ and $\sum_{i=n+1}^{N}(x_i - \bar{x}_{N-n})^2$ are with n-1 and N-n-1 degrees of freedom respectively. Z can be written as

$$Z = \frac{\quad}{\quad + \quad} \qquad (4.17)$$

In this case and are independent. Hence the distribution of Z is a Beta distribution with parameters $\frac{n-1}{2}$ and $\frac{N-n}{2}$.

$$P\left(\frac{L_2}{L_1}\right) = P\left(\frac{L_2^2}{L_1^2}\right) = P\left(Z \frac{n-1}{N-1} \cdot \frac{t_1^2}{t_2^2}\right)$$

$$= \chi(p,q) \qquad (4.18)$$

where $x = \frac{n-1}{N-1} \cdot \frac{t_1^2}{t_2^2}$, $p = \frac{n-1}{2}$, $q = \frac{N-n}{2}$ and

$$\chi(p,q) = \frac{(p+q)}{(p)\ (q)} \qquad (4.19)$$

Table 4 - Giving the distribution function of $\frac{L_2}{L_1}$ for N = 100 and n = 10,20,25.

	$P\left(\frac{L_2}{L_1}\right)$		
	n = 10	n = 20	n = 25
.2	0.0002	-	-
.4	.0011	-	-
.6	.0184	0.0013	0.0003
.8	.1053	.0498	.0325
1.0	.3198	.3579	.3692
1.2	.6184	.8190	.8793
1.4	.8593	.9870	.9968
1.6	.9639	.9999	1.0000
1.8	.9944	1.0000	
2.0	.9995		
2.2	1.0000		

5. STRATIFIED SAMPLING

Let us now consider a case where the population is divided into strata and from each stratum n independent and interpenetrating subsamples have been selected. Let y_{ij} be an unbiased estimate of $_j$, the j th stratum total, from the i th subsample (j = 1,2 .... k;

$i = 1, 2 \ldots n$). The object is to set up confidence interval for $= \sum_{j=1}^{k}$ $_j$, the population total. For this two methods have been suggested and their efficiencies compared. Let us assume that y_{ij} is distributed normally with mean $_j$ and standard deviation $_j$. Then an unbiased estimate of is given by $Y = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} Y_{ij}$. In fact y is distributed normally with mean and variance $\frac{1}{n} \sum_{j=1}^{k}$ ^2_j. The following two estimates of this variance can be considered.

$$(i) \quad s_1^2 = \frac{1}{n(n-1)} \sum_{j=1}^{k} \sum_{i=1}^{n} (Y_{ij} - \bar{Y}_j)^2 \qquad (5.1)$$

and $$(ii) \quad s_2^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} (Y_{i.} - Y)^2 \qquad (5.2)$$

where $\bar{Y}_{.j} = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}$ and $Y_{i.} = \sum_{j=1}^{k} Y_{ij}$.

The variance of these estimates have been compared in section 3. If $_j =$ for all j, then $n(n-1) \frac{s_1^2}{k^2}$ and $n(n-1) \frac{s_2^2}{k^2}$ are distributed as with $k(n-1)$ and $(n-1)$ degrees of freedom. Hence the statistics

$$t_1 = \frac{Y -}{s_1} \qquad \text{and} \qquad t_2 = \frac{Y -}{s_2}$$

will be distributed as Student's t with $(n-1)$ k and $(n-1)$ degrees of freedom respectively. If t and t. are the percentage points of t distribution with $k(n-1)$ and $(n-1)$ degrees of freedom respectively, then the lengths of the confidence intervals based on s_1 and s_2 are given by

$$L_1 = 2t_1 \, s_1 \qquad \text{and} \qquad L_2 = 2t_2 \, s_2$$

If k is fairly large t_1 = 1.96 and the table 1 gives the ratio of $E(L_2)/E(L_1)$ for different values of n.

6. DETERMINATION OF SAMPLE SIZE

Let Y be normally distributed with mean and standard deviation . Suppose a sample of N units is drawn with equal probability. Then the mean $\bar{Y}_N$ based on the N observations is normally distributed with mean and standard error $\frac{}{\sqrt{N}}$. Let s^2 be an unbiased estimator of ^{2} based on a sub-sample of n units,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2.$$

It is well known that the statistic

$$t = \frac{\bar{Y} - }{s/\sqrt{N}}$$

is distributed as Student's t with (n-1) degrees of freedom.

Using the tabulated values of the t-distribution, we can set up a confidence interval for at any specified level of confidence (1 -). That is, if t is the point of t_1 then

$$P \quad \frac{\bar{Y} - }{s/\sqrt{N}} \quad t \quad = 1 - \qquad (6.1)$$

The length L of the confidence interval is given by

$$L = 2 \, t \, \frac{s}{\sqrt{N}} \, . \qquad (6.2)$$

Suppose the sample size is to be so fixed that

$$P \quad L \quad = 1 - \qquad (6.3)$$

where k is a pre-specified quantity and (1-) may be taken as the second level of confidence, the first level of confidence being (1-) in (6.1). It may be noted that P (L k) is a function of the sample size and increases with increase with sample size.

For finding the sample size which would satisfy both the levels of confidences given in (6.1) and (6.3), we may proceed as follows:

$$P \quad L \quad k \quad = 1 -$$

$$\text{i.e.,} \quad P \quad = 1 -$$

$$\text{(i.e.),} \quad P \quad \frac{(n-1)s^2}{2} \quad \frac{k^2}{c^2} \quad \frac{N(n-1)}{4t^2} \quad = 1 - \qquad (6.4)$$

where C is the population coefficient variation and $\frac{(n-1)s^2}{2}$ is a with (n-1) degrees of freedom. Reducing (6.4) to an incomplete function which is already tabulated, we get

$$P \quad L \quad k$$

where

$$p = \frac{n-3}{2}, \quad \text{and} \quad u = \frac{k^2}{c^2} \quad \frac{N\sqrt{n-1}}{4\sqrt{2}\,t^2} \qquad (6.5)$$

For given values of (1-), (1-), c, n and k_2 we can first get the value of u such that

$$I (u_1 p) = 1-$$

and then get the required sample size

$$N = u \frac{c^2}{k^2} \quad \frac{4\sqrt{2}\,t^2}{\sqrt{n-1}} . \qquad (6.6)$$

The values of N for different values of , e, k and n are being calculated. In this chapter only specimen tables giving the value of N for $1 - \quad = 0.95$, $1 - \quad = 0.95$, $e = 1.0$, $k = 0.01(0.01)$ $0.1(0.02)0.2(0.05)0.50(0.10)1.00$ and $n = 2(1)\ 30$, are being given. So far, the procedure of determining the sample size consisted in finding the value of N such that $t\ \frac{e}{\sqrt{N}}$ is equal to a specified value. The proposed procedure given in this chapter is an improvement over the previous procedure, since in the proposed procedure, the variation in the length of the confidence interval is also taken into account.

Table - Showing the sample size required to provide a 95% confidence interval whose length relative to the parameter is less than k with probability 95% when the population coefficient of variation is 1.0.

n \ k	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0;12	0.14	0.16
2	2481 3176	620 3294	275 7019	1550824	992527	689255	506391	387706	306336	248132	172314	126598	96926
3	221 8193	55 4548	24 6466	138637	88728	61616	45269	34659	27385	22182	15404	11317	8665
4	105 5211	26 3804	11 7246	65951	42209	29312	21535	16488	13027	10552	7328	5384	4122
5	73 1487	18 2871	8 1276	45718	29259	20319	14928	11429	9031	7315	5080	3732	2857
6	58 5444	14 6361	6 5049	36590	23418	16262	11948	9148	7228	5854	4066	2987	2287
7	50 2797	12 5699	5 5866	31425	20112	13967	10261	7856	6207	5028	3492	2565	1964
8	44 9770	11 2443	4 9975	28111	17991	12494	9179	7028	5553	4498	3123	2295	1757
9	41 2437	10 3109	4 5826	25777	16497	11457	8417	6444	5092	4124	2864	2104	1611
10	38 4762	9 6191	4 2752	24048	15391	10688	7852	6012	4750	3848	2672	1963	1503
11	36 3539	9 0885	4 0393	22721	14542	10098	7419	5680	4488	3635	2525	1855	1420
12	34 6616	8 6654	3 8513	21664	13865	9628	7074	5416	4279	3466	2407	1768	1354
13	33 2780	8 3195	3 6976	20799	13311	9244	6791	5200	4108	3328	2311	1698	1300
14	32 1425	8 0264	3 5673	20066	12842	8918	6552	5016	3964	3211	2230	1638	1254
15	31 1425	7 7856	3 4603	19464	12457	8651	6356	4866	3845	3114	2163	1589	1217
16	30 2787	7 5697	3 3643	18924	12111	8411	6179	4731	3738	3028	2103	1545	1183
17	29 5683	7 3921	3 2854	18480	11827	8213	6034	4620	3650	2957	2053	1509	1155
18	28 9103	7 2276	3 2122	18069	11564	8031	5900	4517	3569	2891	2008	1475	1129
19	28 3215	7 0804	3 1468	17701	11329	7867	5780	4425	3496	2832	1967	1445	1106
20	27 8057	6 9514	3 0895	17379	11122	7724	5675	4345	3433	2781	1931	1419	1086
21	27 3445	6 8361	3 0303	17090	10938	7596	5581	4273	3376	2734	1899	1395	1068
22	26 9327	6 7332	2 9925	16833	10773	7481	5496	4208	3325	2693	1870	1374	1052

														1843	1354	
23	26	5355	6	6839	2	9484	16585	10614	7371	5415	4146	3276	2654	1819	1336	1037
24	26	1908	6	5477	2	9101	16369	10476	7275	5345	4092	3233	2619	1796	1320	1023
25	25	8648	6	4662	2	8739	16166	10346	7185	5279	4041	3193	2586	1776	1305	1010
26	25	5706	6	3927	2	8412	15982	10228	7103	5218	3995	3157	2557	1756	1290	999
27	25	2910	6	3227	2	8101	15807	10116	7025	5161	3952	3122	2529	1738	1277	988
28	25	0334	6	2584	2	7815	15646	10013	6954	5109	3911	3091	2503	1729	1270	978
29	24	8946	6	2237	2	7661	15559	9958	6915	5081	3890	3073	2489	1705	1253	972
30	24	5526	6	1381	2	7281	15345	9821	6820	5011	3836	3031	2455			959

n \ k	0.18	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.60	0.70	0.80	0.90	1.00
2	76584	62033	39701	27570	20256	15508	12253	9925	6893	5064	3877	3063	2481
3	6846	5545	3549	2465	1811	1386	1095	887	616	452	347	274	222
4	3257	2638	1688	1172	861	660	521	422	293	215	165	130	106
5	2258	1829	1170	813	597	457	361	293	203	149	114	090	73
6	1807	1464	937	650	478	366	289	234	163	120	91	72	59
7	1552	1257	804	559	410	314	248	201	140	103	79	62	50
8	1388	1124	720	500	367	281	222	180	125	92	70	56	45
9	1273	1031	660	458	337	258	204	165	115	84	64	51	41
10	1188	962	616	428	314	240	190	154	107	79	60	47	38
11	1122	909	582	404	297	227	180	145	101	74	57	45	36
12	1070	867	555	385	283	217	171	139	96	71	54	43	35
13	1027	832	532	370	272	208	164	133	92	68	52	41	33
14	991	803	514	357	262	201	158	128	89	66	50	40	32
15	961	779	498	346	254	195	154	125	87	64	49	38	31
16	934	757	484	336	247	189	150	121	84	62	47	37	30
17	913	739	473	329	241	185	146	118	82	60	46	37	30
18	892	723	463	321	236	181	143	116	80	59	45	36	29
19	874	708	453	315	231	177	140	113	79	58	44	35	28
20	858	695	445	309	227	174	137	111	77	57	43	34	28
21	844	684	438	304	223	171	135	109	76	56	43	34	27
22	831	673	431	299	220	168	133	108	75	55	42	33	27

n \ k	0.18	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.60	0.70	0.80	0.90	1.00
23	819	663	424	295	217	166	131	106	74	54	41	33	27
24	808	655	419	291	214	164	129	105	73	53	41	32	26
25	798	647	414	287	211	162	128	104	72	53	40	32	26
26	789	639	409	284	209	160	126	102	71	52	40	32	26
27	781	632	405	281	206	158	125	101	70	52	40	31	-
28	773	626	400	278	204	156	124	100	70	51	39	31	-
29	768	622	398	277	203	156	123	100	69	51	39	31	-
30	758	614	393	273	200	154	121	98	68	50	38	30	-

REFERENCES

1. ASTM Committee (1951) 'Manual on Quality Control of materials', American Society for testing materials.

2. Cadwell, J.H. (1954) 'The probability integral of range from a symmetrical unimodal population' AMS, 25, 803-806.

3. Daly, J.F. (1946) 'The use of sample range in an analogue of Student's t test' AMS, 17, 71-74.

4. Evans, W.D. (1951) 'On the variance of estimates of standard deviation and variance' JASA, 46, 220-224.

5. Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953) 'Sample Survey Methods and Theory' Vol.1, Ch.10, John Wiley & Sons, New York.

6. Jackson, J.E. and Ross, E.L. (1955) 'Extended tables for use with the G-test' JASA, 50, 416-433.

7. Lord, E. (1947) 'The use of range instead of standard deviation in the t test' Biometrika, 34, 41-6%.

8. Mahalanobis, P.C. and others (1934) 'Tables of random samples from a normal population, 'Sankhya', 1, 289.

9. Nair, K.R. (1940) 'Tables of confidence interval for the median in samples from any continuing population', Sankhya, 4, 551-558.

10. Noether, G.E. (1955) 'Use of the range instead of the standard deviation', JASA, 50, 1040-1055.

11. Patnaik, P.B. (1950) 'The use of mean range as an estimate of variance in statistical tests' Biometrika, 37, 78-87.

12. Pearson, E.S. (1932) 'The percentage limits to the distribution of range in samples from a normal population' Biometrika, 24, 404-417.

13. Pearson, E.S. and Maines, J. (1935) 'The use of range in place of standard deviation in small samples' Supplement to JRSS, 2, 83-98.

14. Stevens, W.L. (1948) 'Control by Gauging' JRSS (Series B), 10

15. Thompson, W.R. (1936) 'On confidence ranges for the median and other expectation distributions for populations of unknown distribution form', AMS, 7, 122-128.

Chapter 7

SELF-WEIGHTING DESIGN AT FIELD AND TABULATION STAGES

1. INTRODUCTION

A sampling design is said to be self-weighting if the weights to be given to the values of the selected sampling units are the same. The technique of making a design self-weighting at field stage (that is, selection of units for enquiring in such a way as to make the weights the same) is used in many of the surveys because of the considerable saving in tabulation time. Further it is also believed that for many of the situations commonly met with in practice, a self-weighting design would be more efficient than a non-self-weighting design. Hansen, Hurwitz and Madow (1953) have considered this technique in detail in their book.

In this chapter, it is proposed to present the technique of making a design self-weighting at field stage and to discuss its implications in large scale surveys. The present author has investigated the question of making a design self-weighting at tabulation stage in collaboration with Mr. V.K. Sethi and the results of this investigation are presented in sections 8 and 10 of this chapter. A number of procedures of making the design self-weighting at tabulation stage have been considered. The efficiencies of these procedures are studied empirically and the results of this study are also given in this chapter.

2. SELF-WEIGHTING DESIGN

In sample surveys the estimators commonly used for estimating the population total of a variate are of the form

$$\sum w_i y_i$$

where y_i is the value of the ith selected ultimate sampling unit with w_i as its weight and $\sum$ is over all the sample observations. The weight w_i depends on the selection and estimation procedures. These weights are known as 'multipliers', 'inflation factors' or 'raising factors' since they are used to inflate the sample observations to get an estimate of the population total. For instance, in case of equal probability sampling, with or without replacement or systematically the multiplier is common for all the selected units and is given by $\frac{N}{n}$, the inverse of the sampling fraction, whereas in case of pps sampling with replacement the multiplier differs from unit to unit and for the ith selected unit it is $\frac{1}{n}\frac{1}{p_i}$ where p_i is the probability of selection. In a two stage design the multiplier is of the form $\frac{1}{n}\frac{1}{p_i}\frac{1}{m_i}\frac{1}{p_{ij}}$ where m_i is the number of second stage units selected in the ith first stage unit and p_{ij} is the probability of selection of the jth second stage unit in the ith first stage unit.

In large scale surveys where a number of characteristic are to be estimated the calculation at estimation stage becomes difficult and time consuming if the multiplier varies from unit to unit. Hence from the consideration of ease at tabulation stage it is very

desirable to have a sampling design which gives rise to a single common multiplier for all the sampled units. Such a sampling design is called a 'self-weighting design' because little effort is necessary in this case in weighting the sample observations. Further it is believed that a self-weighting design which utilises all the available information is more efficient than other designs since it gives the same chance of selection to the ultimate sampling units.

It is possible to make a design self-weighting either at the field stage or at the tabulation stage. In case of the former, selection of units at the ultimate stage in a multi-stage design is so arranged as to make the design self-weighting whereas in the latter case some technique is devised at the estimation stage to make the sampling units have the same multiplier. Usually it would be desirable to adopt the former procedure in preference to the latter. The latter procedure is to be adopted only if it is not possible to make the design self-weighting at field stage due to some operational considerations.

Though it would be ideal to have one common multiplier for all the selected units, there would be considerable saving at tabulation stage even if two or more common multipliers instead of one are to be used provided the number of such common multipliers is fairly small. In the latter case the design may be said to be partially self weighting. It may be mentioned that the latter

situation where a small number of common multipliers are to be used instead of one such is likely to arise more often than the former situation due to the restrictions usually imposed on the design by operational considerations.

3. STRATIFIED UNISTAGE SAMPLING

In a stratified simple random sampling with or without replacement or systematically, the estimate of population total Y is given by

$$\hat{Y} = \sum_{s=1}^{k} \frac{N_s}{n_s} \sum_{i=1}^{n_s} y_{si} \qquad (3.1)$$

where N_s and n_s are respectively the number of units and sample size in the sth stratum and y_{si} is the value of the ith selected unit in that stratum. Hence the multiplier is $\frac{N_s}{n_s}$. If the total sample size n is fixed, the design can be made self-weighting by adopting proportional allocation to the strata, for in that case

$$\frac{N_s}{n_s} = \frac{N}{n} . \qquad (3.2)$$

This shows that the common multiplier is the inverse of the overall sampling fraction.

If no information is available about the variation in the strata, proportional allocation which makes the design self-weighting is likely to be the most efficient. Comparing the variances of the estimator in case of equal allocation and proportional allocation under the assumption of equal probability sampling with replacement

in the strata we get,

$$V_{eq} - V_{prop.} = \frac{1}{n} \sum_{s=1}^{k} (k N_s - N) N_s \sigma_s^2 \qquad (3.3)$$

where σ_s^2 is the variance in the sth stratum. This shows that $V_{prop.}$ is being to be less than V_{eq} since the variance in larger strata can be expected to be more than in smaller strata.

It is to be noted that use of self-weighting design imposes restrictions on the design. For instance in the above case, the allocation has to be proportional even if we have knowledge about the strata variances since use of optimum allocation makes the design non-self-weighting. In such a case a decision regarding use of proportional allocation and optimum allocation is to be arrived at after considering the relative magnitudes of the gain in tabulation cost and of the loss in precision of the estimator. Further, adoption of self-weighting design makes the work-load differ from stratum to stratum which may not be desirable from operational considerations.

If the numbers of units in the strata are small, then there would be rounding off errors in having proportional allocation. This can be avoided in a convenient way by selecting the units in the different strata systematically with interval $\frac{N}{n}$ and random strata from 1 to $\frac{N}{n}$. It may be noted that stratified proportional allocation sampling can be achieved if the sample is selected systematically from the whole population after arranging the units

stratum-wise. Approximately equal work-load in the different strata can be achieved by suitably changing the sampling intervals to be used in them. But when this procedure is adopted the design can only be partially self-weighting and not perfectly self-weighting. In other words there would not be one common multiplier but a small number of multipliers would have to be used.

The following example would clarify the points mentioned above. Suppose there are 6 strata and 162 units in the population. Let the sample size be 20.

Table 1 : Showing the values of sampling interval and the expected values of the sample size in the different strata for two schemes.

scheme	stratum s	1	2	3	4	5	6
	no. of units N_s	52	14	25	28	12	31
i) self-weighting but unequal work-load	sampling interval I_s	8.1	8.1	8.1	8.1	8.1	8.1
	$E(n_s)$	6.4	1.7	3.1	3.5	1.5	3.8
ii) partially self-weighting with equal work-load of about 3 units per stratum	modified I_s	16.2	4.05	8.1	8.1	4.05	8.1
	modified $E(n_s)$	3.2	3.4	3.1	3.5	3.0	3.8

From the above table it may be seen that in the first case the allocation is proportional and there is one common multiplier

though the work-load varies from stratum to stratum where as in the second case there are 3 different multipliers to be used and the allocation is not proportional but there is little variation in work-load. The former design is likely to be more efficient than the latter. But the latter may have to be adopted if there is considerable advantage in making the work-load approximately the same in the different strata.

If the sampling interval is not an integer, it is possible to select a systematic sample with this interval by associating a suitable number of numbers with the units. For instance if the interval is 8.1, the interval 81 can be used after associating 10 numbers with each unit. For the sake of simplicity we may round off 8.1 to 8 or 9 with probabilities 0.9 and 0.1 respectively.

In a stratified design where the units in the strata are selected with probability proportional to a given measure of size x with replacement, the estimator of the population total Y is given by

$$Y = \sum_{s=1}^{k} \frac{X_s}{n_s} \sum_{i=1}^{n_s} \frac{y_{si}}{x_{si}} \qquad (3.4)$$

with the usual notation. If the ratio $\frac{y_{si}}{x_{si}}$ can be readily observed in the field or can be reported by the investigator without much difficulty, the design would become self-weighting if the allocation is done in proportion to X_s, the total size of the sth stratum. For instance, in case of a crop survey y and x

may stand for area under a particular crop and geographical area respectively in which case the ratio $\frac{y_{si}}{x_{si}}$ is the proportion of the area under the crop in the sampling unit which may be a plot or field and can be easily reported. In this case the estimator would be

$$\hat{Y} = \frac{X}{n} \sum_{s=1}^{k} \sum_{i=1}^{n_s} p_{si} \qquad (3.5)$$

where p_{si} is the proportion of the area under the crop in the ith selected unit of the sth stratum, X is the total geographical area and n the total sample size. In a socio-economic survey where the single unit is a household y and x may stand for expenditure on a given item and household size. Then the ratio to be reported is per capita expenditure for a household.

4. STRATIFIED TWO-STAGE SAMPLING

In a stratified two stage design where the first stage units are selected with probability proportional to a given measure of size with replacement and the second stage units are selected systematically an unbiased estimator of the population total Y is given by

$$\hat{Y} = \sum_{s=1}^{k} \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{1}{p_{si}} \frac{M_{si}}{m_{si}} \sum_{j=1}^{m_{si}} y_{sij} \qquad (4.1)$$

where the subscripts s, i and j stand respectively for sth stratum, ith selected forst stage unit and jth selected second stage unit,

n is the number of first stage units selected M and m are respectively the number of second stage units in the population and sample respectively and y is the value of the characteristic under consideration. Hence the multiplier is

$$\frac{1}{n_s} \quad \frac{1}{p_{si}} \quad \frac{M_{si}}{m_{si}} \qquad (4.2)$$

It may be noted that even if the first stage units are selected pps systematically and second stage units are selected with equal probability with or without replacement the multiplies remains the same as that given above. For instance, in a socio-economic survey one may select the village with probability proportional to population with replacement or systematically and in the second stage, select the households with equal probability with or without replacement or systematically. Similarly in case of crop survey villages may be selected in the first stage with probability proportional to the survey numbers with replacement or systematically and in the second stage survey numbers may be selected with equal probability with or without replacement or systematically.

To make the design self-weighting we have to fix m_{si} the number of second stage units to be selected in the ith first stage unit in the sth stratum such that the above multiplier is constant C. It may be seen that for any given constant C the design becomes self-weighting if

$$m_{si} = \frac{1}{C} \frac{M_{si}}{n_s \, p_{si}} \qquad (4.3)$$

From this it is clear that if C is made large the sample size in terms of the number of second stage units would be reduced since m_{si} would be small and similarly if C is taken to be small then the number of second stage units in the sample would be larger than the requirement. Hence C has to be determined so as to get the required sample size on the average. If a sample of m second stage units is to be selected on the average from a selected first stage unit, we get,

$$\sum_{s=1}^{k} \sum_{i=1}^{n_s} m_{si} = \frac{1}{C} \sum_{s=1}^{k} \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{M_{si}}{p_{si}} = nm \qquad (4.4)$$

and hence

$$C = \frac{\sum_{s=1}^{k} \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{M_{si}}{p_{si}}}{nm} \qquad (4.5)$$

This shows that the constant C should be taken as the inverse of the over all sampling fraction for the second stage units since the numerator in (4.5) is an unbiased estimate of the total number of second stage units. Once C is determined, the design can be made self-weighting by selecting the required number of second stage units from the selected first stage units using the appropriate sampling intervals. If the current values of M_{si} are not available, the values of M_{si} for a previous period may be used as an approximation for determining C.

The sampling interval to be used in the ith selected first stage unit in the sth stratum is given by

$$I_{si} = \frac{M_{si}}{m_{si}} = C\, n_s\, p_{si} \qquad (4.6)$$

In this case the investigator would be supplied with the interval and the random start from 1 to I_{si} (and not from 1 to M_{si}). The sample of second stage units would be selected with the specified random start and sampling interval. It may be noted that for this procedure it is not necessary to know the M_{si} in advance. Of course the sample size actually obtained would depend on the M_{si}'s. If I_{si} is not an integer, it may be suitably rounded-off as mentioned earlier.

It may be seen that in case of stratified two stage design where the first stage units are selected with pps and the second stage units are selected with equal probability, the design can be made self-weighting with equal work-load in the selected first stage units by allocating the sample size to the strata in proportion to the total stratum-size and selecting the same number of second stage units from each selected first stage unit. With this type of design, the work-load in a stratum can be equalized by forming strata with approximately the same total size.

The above discussion shows that it is possible to make a given design self-weighting by suitably determining the sampling interval to be used in the selected ultimate stage units and to ensure at the

same time the required sample size on the average. Here also two or more common multipliers may be used if found necessary from field operational considerations.

5. SELF-WEIGHTING DESIGN AT TABULATION STAGE

One of the main difficulties in making the sampling design self-weighting at the field stage is that the work-load in the penultimate stage units becomes unequal which is not desirable from the point of view of administrative consideration in large scale surveys. The methods which make a design self-weighting at field stage ensuring at the same time equal work-load within penultimate stage units impose rather severe restrictions on the allocation and selection procedures. Because of these considerations it may not be always advisable or feasible to adopt a self-weighting design at field stage.

Suppose N units have been selected from a population according to some specified non-self-weighting design. Let w_i and y_i be the multiplier and the value of the characteristic of the ith selected sampling unit respectively. Then the estimate of the population total Y for this characteristic is given by $\sum_{i=1}^{n} w_i y_i$ (=Y). The problem here is to find a technique by which the set of multipliers can be replaced by a smaller number of multipliers such that the estimate remains unbiased and the increase in variance is not much or if the estimate becomes biased the bias is negligible.

6. ROUNDING-OFF TECHNIQUES

The following five procedures of making the design self-weighting at the tabulation stage are considered:

(i) rounding-off to the nearest hundred, thousand or ten thousand;

(ii) substituting each multiplier by the mean of the multipliers;

(iii) optimum set of rounded-off multipliers;

(iv) sub-sampling with probability proportional to the multipliers with replacement;

and (v) sub-sampling with probability proportional to the multipliers systematically.

The procedures (i) and (ii) give biased estimates with possibly decrease in the variance of the estimate under certain circumstances, while the procedures (iii), (iv) and (v) give unbiased estimates with some increase in the variance. It may be mentioned that in some cases the procedures (i) and (v) are being adopted in some actual surveys to expedite the tabulations.

Procedure (i) : Suppose we have a series of four digit multipliers. The usual procedure of decreasing the number of multipliers is to round-off the multipliers to the nearest thousand. Thus the N four digit multipliers are replaced by a set of ten rounded-off multipliers. The estimate obtained by using this method will obviously be biased. The bias depends much on the multipliers as well as the

values of the characteristics In practice it is difficult to get an idea of the sign and magnitude of the bias unless one works out the biased estimate as well as the estimate using the actual multipliers.

Procedure (ii) : If this method is adopted, the estimate is given by the product of the mean of the multipliers and the sum of the values of the characteristic. This estimate is also biased, but the bias will be negligible if in the sample, the co-variance of the multiplier and the value of the selected unit is very small. For instance in a particular per capita expenditure class, it is felt, the expenditure on cereals, food and total expenditure of a household might not depend much on its multiplier. In other words, the expenditure on cereals, food and total expenditure of a household, belonging to a particular per capita expenditure class are expected to depend on the household size, the geographical, the climatic and the economic conditions of the locality and not so much on whether the household is in a large village or a small village. If only the estimates of expenditure on cereals and such items are required, the procedure would be to classify the household according to the per capita expenditure class and then find the biased estimates separately for all the classes. The sum of these biased estimates will be the estimate of the population total and the bias can be expected to be small. It may be noted that in this case the number of multipliers will be equal to the number of per capita expenditure classes.

In this connection, it should also be noted that this method is not to be used indiscriminately as there may be characteristics which are related to the multipliers. This method can be used only if we are fairly certain that the value of the item in which we are interested is uncorrelated with the multiplier.

Procedure (iii) : A general solution will be to round-off each of the multipliers to a certain number of weights which may be called rounded-off multipliers with such probabilities that the expected value is the original multiplier. As these weights are at our choice, we can choose them such that the increase in variance is minimised. Further a desired increase in the variance at the tabulation stage can be got by taking a sufficient number of rounded-off multipliers in the optimum fashion.

As the optimum solution for a specified number of rounded-off multipliers depends on the value of the characteristic in question, in practice it is not possible to get the optimum solution. Of course some method which will give us a solution near about the optimum can be devised. Another difficulty is that the determination of the approximate optimum rounded-off multipliers becomes more and more difficult as the number of such multipliers is sought to be increased. This procedure is considered in detail in section 8.

Procedure (iv) : The method consists in taking a sub-sample of n' units from the field sample of size n with probability proportional to the multipliers. The sum of the values in the sub-sample

multiplied by a constant (the ratio of sum of multipliers to n') gives an unbiased estimate of $\sum_{i=1}^{n} w_i y_i$. The increase in variance at the tabulation stage is given by

$$E_f \frac{1}{n'} \left(\sum_{i=1}^{n} w_i \right) \sum_{i=1}^{n} w_i y_i^2 - \left(\sum_{i=1}^{n} w_i y_i \right)^2 \qquad (6.1)$$

where E_f is the expected value over the field sample. It can be seen that this method results in rounding-off of each of the multipliers to one of the set of multipliers $\left(j \frac{\sum w_i}{n'}\right)$, $j = 0,1,2,\ldots n'$ with certain probabilities such that the estimate remains unbiased and the sum of the round-off multipliers in the sub-sample is equal to the sum of the actual multipliers.

If instead of sub-sampling, we round-off each of the multipliers to one of the set of multipliers $\left(j \frac{\sum w_i}{n'}\right)$ $j = 0,1,2, \ldots n'$ with probabilities $\binom{n'}{j} \left(\frac{w_i}{\sum w_i}\right)^j \left(1 - \frac{w_i}{\sum w_i}\right)^{n'-j}$ then the estimate is unbiased and the increase in variance is given by

$$E_f \frac{1}{n} \sum_{i}^{n} w_i \sum_{i}^{n} w_i y_i^2 - \sum_{i}^{n} w_i^2 y_i^2 \qquad (6.2)$$

Comparing (6.1) and (6.2) we see that the expression (6.2) is greater than (6.1) since $w_i y_i$'s are positive in general. Hence sub-sampling with probability proportional to the multipliers with replacement is more efficient than the corresponding randomising method.

Procedure (v) : In this method the units in the field sample are arranged in a certain manner, the multipliers are cumulated and a systematic sample is drawn with the interval $\sum w_i/n'$. As in the

case of pps with replacement, the estimate is given by the product of the sum of the values in the sub-sample and a constant which is the interval. An expression for the increase in variance is difficult to find in this case. This method results in rounding-off of each of the multipliers (say w_i) to either $\frac{n' w_i}{w_i} \frac{w_i}{n'}$ or $\frac{n' w_i}{w_i} + 1 \quad \frac{w_i}{n'}$ with certain probabilities such that the estimate remains unbiased and the sum of the rounded-off multipliers in the sub-sample is equal to the sum of the multipliers in the field sample.

A sort of 'without replacement' element is present in this pps systematic selection. For when the units are arranged at random and are equal, then the above procedure amounts to taking a simple random without replacement, while in that case the pps with replacement procedure will amount to taking a simple random sample with replacement. Further drawing of a sub-sample in the case of pps systematic is likely to take less time than in the case of pps with replacement as in the former case we need not refer to random number tables more than once. Also by properly arranging the units and devising a suitable balancing procedure the estimates can be improved upon in pps systematic method.

7. EFFICIENCY OF SELF-WEIGHTING DESIGN

The following notation is used:

N : number of villages in the population

n : " " " " " sample

N_i : " " households in the ith village

n_i : " " " " " " " included in the sample

p_i : probability of selecting the ith village

h_i : average household size in the ith village

h : " " " " " population

$M_o = \sum_i N_i$

y_{ij} : value of the characteristic for the jth household of the ith village

$\bar{y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$

$\bar{y} = \frac{\sum_i \sum_j y_{ij}}{M_o}$

Suppose the sampling design is a two stage one where the first stage units are selected with ppp with replacement and the second stage units are selected with equal probability. The following three variations of the above design are considered:

(i) n villages are selected with ppp with replacement and m households are selected from each selected village with equal probability without replacement.

(ii) n villages are selected with ppp with replacement and with (where $= \frac{M_o}{mn}$) as the sampling interval, the households are selected after arranging them at random.

(iii) The sample is selected as in (i). Then the weights are rounded-off to two points and at random where and

The estimates and the variances in the above three cases are given by

where stands for summation over all households whose weights are rounded-off to and for summation over the rest of the sample households.

where is the coefficient of variation of the characteristic in the population in the ith village and

It is difficult to compare the above variances in general and extensive empirical studies may be necessary before one method is preferred to another in particular cases.

Empirical study

Data used : 1951 census data for Siruguppa tehsil of Bellary District.

Characteristics : i) Number of males in M/L class I (y_1) 19901

ii) " " " " " " (y_2) 5549

= 0.70 ; = 1.25

	characteristics (i)				characteristic (ii)			
	c = 1	m = 1	c = 2	m = 1	c = 1	m = 1	c = 2	m = 1
$\frac{v_2 - v_1}{v_1} \times 100$	16.31		9.26		10.82		7.77	
$\frac{v_3 - v_1}{v_1} \times 100$	13.46		3.46		8.74		2.88	

where $c_i = c$ for all i.

8. RANDOMISED ROUNDED-OFF MULTIPLIERS

Suppose y_i and a_i (i = 1, 2, ..., N) denote the characteristic value of the ith sample unit and its corresponding multiplier such that $\sum_{i=1}^{N} a_i y_i$ is an unbiased estimate of the population total Y. The problem is to find a set of n rounded-off multipliers

$$b_i \quad , \quad i = 1, 2, \ldots, n,$$

(n being considerably smaller than N) such that

$$E \sum_{i=1}^{N} r_i y_i = \sum_{i=1}^{N} a_i y_i \qquad (8.1)$$

and $V \sum_{i=1}^{N} r_i y_i$ is minimum

where r_i is a random variable taking the values $\{b_i\}$ with certain probabilities, E stands for the conditional expectation given the sample of N units and V stands for the conditional variance given the sample of N units. Further a balance is to be struck between the increase in variance and the amount of labour involved in obtaining and using these rounded-off multipliers.

A necessary and sufficient condition for the equation (8.1) to hold for all values of y is that $E(r_i) = a_i$ for all i. Suppose the set of n rounded-off multipliers b_i is given and the range of this set includes the range of the original multipliers. $E(r_i)$ will be equal to a_i and $V(r_i)$ will be minimum if r_i takes only the values b_k and b_{k+1} nearest to a_i on both sides of it (b_k a_i b_{k+1}) with probabilities

$$\frac{b_{k+1} - a_i}{b_{k+1} - b_k} \quad \text{and} \quad \frac{a_i - b_k}{b_{k+1} - b_k} \tag{8.2}$$

respectively.

With this procedure of allocation of the rounded-off multipliers to the original multipliers, the values of the rounded-off multipliers which would minimise the increase in variance of the estimate are given by

$$b_1 = a_1 \quad , \quad b_n = a_N \quad \text{and}$$

$$b_j < \sum_{a_i < b_{j+1}} (y_i^2) < \frac{\sum_{b_{j-1} < a_i < b_{j+1}} (a_i - b_{j-1}) y_i^2}{(b_{j+1} - b_{j-1})} < \sum_{b_j < a_i < b_{j+1}} (y_i^2) \quad (j=2,3,\ldots, n-1) \quad \ldots (8.3)$$

where a_1 and a_n are respectively the smallest and the largest multipliers. In practice, all the units having original multipliers between b_{j-1} and b_{j+1} may be arranged in decreasing order of the multipliers. Then the cumulated sums of y_i^2 for an auxiliary or key characteristic are to be determined from the top ajd the multiplier of that unit where the cumulated total is equal to or just greater than the value of the middle expression in (8.3) is to be taken as the value of b_j. The actual solutions for b_j $j = 2, 3, \ldots (n-1)$ are difficult to obtain if n is greater than 3. When n is greater than 3, the following iteration process is suggested. Suppose n is 4. First the optimum set of three multipliers b_1, b_2, b_4 is determined as indicated above. b_1 and b_4 are the smallest and largest of the original multipliers respectively and b_2 is the optimum rounded-off multiplier between b_1 and b_4 obtained from (8.7). Using (8.3) the optimum rounded-off multiplier b_3 between b_2 and b_4 is determined. Then the optimum rounded-off multiplier b'_2 between b_1 and b_3 is obtained. Between b'_2 and b_4 an optimum rounded-off multiplier b'_3 is found out. This process is repeated till two successive values of the second and third rounded-off multipliers agree. This process can be extended to the case of n greater than 4.

The criterion for the replacement of the multipliers by a smaller number of rounded-off multipliers should be the increase in the variance

of the estimate and this increase is just the expected value of the variance of $\sum_{i=1}^{N} r_i y_i$ over the tabulation stage. Hence the variance of $\sum_{i=1}^{N} (r_i y_i)$ over the tabulation stage is an unbiased estimate of the increase in variance of the estimate.

The following practical method would help in finding the minimum number of rounded-off multipliers required to achieve the desired precision. Since the rounded-off multipliers may be determined one by an iterative process, we may start with n = 3. For n = 3, the value of the second rounded-off multiplier is found out so as to minimise the varance using (33). This together with the largest and the smallest multiplers in the sample would constitute the set of three rounded-off multipliers. The estimate of the increase in variance is found out from this sample as mentioned above. If this is greater than the value decided upon, one more rounded-off multiplier is to be taken. This will be chosen in that part which contributes more to the variance than the other, in such a manner that this contribution to the variance is minimised. This process is continued till the desired variance is achieved.

9. ILLUSTRATION (I)

To try out the technique developed in section 8, the data on consumer expenditure on cereals collected in a large scale survey in Madhya Pradesh were used. Here a_i and y_i stand for the actual multiplier and expenditure on cereals of the ith sample unit.

The a_i's varied from 20961 to 78993. Hence these two numbers were taken as the two extreme rounded-off multipliers b_1 and b_3

respectively. For finding a suitable value for b_2, a sub-sample of size 30 was taken from the sample with equal probability. Using the technique developed in section 3, the value of b_2 was found to be 57600. The increase in the variance of the estimate of total expenditure on cereals was calculated and was found to be 6.67 percent of the estimate of variance of $\sum_{i=1}^{N} a_i y_i$. This meant only 3.28 percent increase in the standard error of the estimate and the same amount of increase in the coefficient of variation.

To test the utility of the additional multiplier b_2, the increase in variance was calculated on the basis of the two multipliers b_1 and b_3. This turned out to be about 170 times the increase in variance when three multipliers were used. This demonstrates that considerable reduction in the increase in variance has been achieved by the additional multiplier b_2.

10. ILLUSTRATION (II)*

As it is not possible, in general, to compare the efficiencies of the procedures suggested in section 6 to reduce the cost at the tabulation stage, an empirical study was conducted to assess their merits and demerits. For this purpose the data on consumer expenditure statistics collected in a large scale sample survey in Uttar Pradesh were used. The object of this study was to cimpare the efficiencies of the procedures (ii), (iv) and (v) as well as their practicability in large scale operations, in estimating (a) expenditure on cereals, (b) expenditure on food, and (c) total expenditure by per capita expenditure classes.

The design of the survey was a stratified three stage one with tehsils as first stage units, villages as second stage units and households as third stage units. From each stratum two tehsils were selected with probability proportional to population (ppp) with replacement. From each selected village about five households were selected systematically with a random start for the Consumer Expenditure Enquiry. It is to be noted that for each stratum total we get two independent estimates one from each of the two tehsils selected from that stratum. The sample households belonging to the tehsils selected first will be considered as belonging to field sample 1(F_1) and the other sample households will constitute field sample 2(F_2).

The sample households belonging to each of the field samples were stratified into fourteen classes on the basis of their per capita total expenditure. The classes considered given below:-

* Part of National Sample Survey Working Paper No. 6 'Self-weighting design at tabulation stage' by Murthy, M.N. and Sethi, V.K.

No. of the class	1	2	3	4	5	6	7	8	9	10	11	12	13	14
per capita expenditure in Rs.)	0-5	5-8	8-11	11-13	13-15	15-18	18-21	21-24	24-28	28-34	34-43	43-55	55-75	75-

In each per capita expenditure class the households were arranged according to the village, tehsil and stratum to which they belonged.

As the per capita expenditure, as such, was not given in the schedule we had to classify the household by making use of the total expenditure and the household size. A method is given below by which the households can be classified into the per capita expenditure classes without actually dividing the total expenditure by the household size. The computers were given copies of a table giving the maximum total expenditure corresponding to the different per capita expenditure classes and household sizes.

For procedures (iv) and (v) from each per capita expenditure class a sample of size equal to the number of households in that class was selected to estimate the total expenditure on cereals, food and total expenditure. This sample size was taken in each class because the comparison of the efficiencies of procedures (iv) and (v) with that of procedure (ii) becomes easy. The variances of the estimates of $\sum_{i=1}^{N} a_i y_i$ got by using procedures (iv) and (v) were calculated assuming the field sample to be the population. These variances can be considered as the estimates of the increase in the variances of the estimate

of population totals of the actual population. The bias of the estimate based on procedure (ii) was also calculated.

A glance at columns (3) and (4) of Tables 2, 3 and 4 show that the pps systematic estimate is much better than the pps with replacement estimate in all the cases. It is possible to improve upon the pps systematic estimate by suitable arrangement of sampling units. So if one is fairly sure that the arrangement of the sampling unit is good, then the pps systematic method will be superior to that of pps with replacement. Of course in the case of pps systematic sampling, the variance cannot be estimated from one sample. This difficulty can be overcome by taking two or more samples with independent random starts.

Table 2 : Showing the precision of different methods of estimation at tabulation stage - (expenditure on cereals).

per capita expenditure class	sub-sample 1				sub-sample 2			
	no. of sample house-holds	bias $\frac{\sum a_i y_i}{\sum a_i y_i}$ x 100 (ii)	c.v.(./.) (iv)	c.v. (./.) (v)	no. of sample house holds	bias $\frac{\sum a_i y_i}{\sum a_i y_i}$ x100 (ii)	c.v. (./.) (iv)	c.v. (./.) (v)
(0)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	4	27.41	43.38	15.47	9	2.47	15.00	7.82
2	18	-8.62	15.44	3.15	19	-1.67	16.24	4.70
3	55	-2.84	7.16	0.84	43	0.76	8.78	4.86
4	31	1.15	8.74	1.21	41	-2.60	30.67	5.62
5	38	-3.07	7.97	1.00	45	-1.98	7.69	3.12
6	58	-1.89	7.60	1.11	52	-4.10	7.26	2.52
7	33	-4.70	8.01	1.88	38	-5.48	11.60	4.64
8	33	-9.73	10.40	1.63	24	-0.53	19.46	5.58
9	23	6.49	11.81	2.83	21	-0.31	15.77	13.06
10	19	-5.61	17.44	2.52	22	-0.32	15.44	7.26
11	24	2.74	13.06	1.72	15	4.11	22.29	13.72
12	10	-26.96	24.28	3.17	5	-4.39	34.86	12.50
13	8	1.41	16.64	2.43	5	-2.91	21.29	11.00
14	4	-4.95	20.31	3.53	2	1.29	33.38	7.96
all	358	-3.36	3.10	1.03	341	-2.36	5.79	1.94

c.v. : coefficient of variation

(ii) substitution by mean of the multipliers

(iv) pps with replacement

(v) pps systematic

Table 3 : Showing the precision of different methods of estimation at tabulation stage - (expenditure on food).

per capita expenditure	sub-sample 1				sub-sample 2			
	no. of sample house holds	$\frac{\text{bias}}{\Sigma a_i y_i}$ x 100 (ii)	c.v.(%) (iv)	c.v. (%) (v)	no. of sample house holds	$\frac{\text{bias}}{\Sigma a_i y_i}$ x 100 (ii)	c.v. (%) (iv)	C.V. (%) (v)
(0)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	4	33.70	47.17	18.07	9	1.28	18.32	9.05
2	18	-5.25	12.87	1.64	19	-0138	10.84	2.04
3	55	-0.46	5.76	0.92	43	1.96	8.61	3.82
4	31	0.18	7.25	0.95	41	-0.31	24.46	4.52
5	38	-0.05	6.25	0.91	45	0.93	6.88	4.54
6	58	1.97	6.36	0.93	52	-2.24	6.38	2.74
7	33	-3.15	7.52	1.69	38	-0.37	8.99	3.88
8	33	-9.87	9.68	1.57	24	-2.00	16.84	6.04
9	23	11.86	10.95	3.63	21	3.86	16.88	6.06
10	19	-4.37	12.24	2.41	22	2.10	11.93	4.68
11	24	-2.08	10.71	2.51	15	5.12	23.17	16.04
12	10	-37.53	30.04	2.48	5	-3.44	21.92	6.98
13	8	-5.61	19.99	3.40	5	-4.19	27.07	10.60
14	4	1.45	18.23	2.50	2	1.00	24.45	6.22
all	358	-2.96	3.07	1.16	341	-0.00	4.72	1.66

c.v. : coefficient of variation

(ii) : substitution by mean of multiplier

(iv) : pps with replacement

(v) : pps systematic

Table 4 : Showing the precision of different methods of estimation at the tabulation stage (total consumer expenditure).

per capita expenditure	sub-sample 1				sub-sample 2			
	no. of sample house holds	$\frac{\text{bias}}{\sum a_i y_i}$ x 100 (ii)	c.v. (%) (iv)	c.v. (%) (iv)	no. of sample house holds	$\frac{\text{bias}}{\sum a_i y_i}$ x 100	c.v. (%) (iv)	c.v. (%) (v)
(0)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	4	33.33	46.97	17.90	9	2.03	20.47	10.62
2	18	-5.04	12.31	1.85	19	-0.42	10.84	2.42
3	55	-0.29	5.40	0.85	43	-1.68	9.62	3.90
4	31	0.42	7.15	0.85	41	-0.28	23.31	12.16
5	38	2.38	6.58	0.97	45	1.19	7.05	5.22
6	58	2.40	6.55	1.05	52	-3.68	6.66	2.50
7	33	-3.51	7.50	1.87	38	-1.88	7.75	3.02
8	33	-8.72	9.93	1.64	24	-1.78	16.65	6.24
9	23	3.30	11.91	2.44	21	12.44	16.22	7.52
10	19	-2.52	11.23	2.08	22	1.69	11.29	6.14
11	24	2.84	11.20	3.18	15	2.52	21.79	11.10
12	10	-30.33	25.46	3.19	5	-5.14	29.41	10.30
13	8	-1.34	20.46	4.30	5	-6.78	48.75	23.64
14	4	13.17	36.86	10.66	2	1.43	35.22	8.86
all	358	-1.63	3.61	1.62	341	0.13	4.89	2.34

c.v. : coefficient of variation

(ii) : substitution by mean of the multipliers

(iv) : pps with replacement

(v) : pps systematic

The comparison of the procedure (ii) with that of pps with replacement shows that the former is to be preferred. But this method has some pitfalls. This will be a good method, as has been already pointed out, only if the correlation coefficient between the multipliers and the variate values is very nearly equal to zero. The systematic pps estimate compared favourably with this biased estimate. Since the pps systematic method has the added advantage of being unbiased, one is justified in preferring that to the biased one.

The difference between the field sample estimates gives us an idea of the variability of the estimate. It may be noted that compared to the variability of the estimate, the increase in variance at the tabulation stage is negligible.

Time Study

To study the relative merits and demerits of the procedures (iv) and (v) from the point of view of operational convenience, time records were kept for the different operations involved in applying the methods (iv) and (v) in practice. The results of this study are given below.

Time Record

Data : Consumer expenditure schedules for the sample households (Rural) in Uttar Pradesh

No. of sample villages : 146
No. of sample households : 699
Characteristic considered : (i) expenditure on cereals,
(ii) expenditure on food, and
(iii) total expenditure.

Table 5 - Showing the time taken for different stages of a tabulation work.

serial number	description of the operation	time taken in hours
1	arrangement of schedules	2
2	classification of households (according to per capita expenditure)	8
3	extracting the information	10
4	copying the village multipliers	2
5	finding of the household multipliers and writing them down	5
6	verification and checking	6
7	getting the cumulants	3
8	getting the usual weighted estimates by cross-multiplication and additions	20
9	getting the systematic pps tabulation sample	1
10	addition and getting the estimates	6
11	selection of pps with replacement tabulation sample	3
12	additions and getting the estimates	6

From the time record we find

(i) time taken for getting the usual weighted estimates involving operations 1 to 6 and 8 is ..... 53 hours

(ii) time taken for getting the systematic pps estimates involving operations 1 to 7, 9 and 10 is ..... 43 hours

and (iii) time taken for getting the pps with replacement estimates involving operations 1 to 7, 11 and 12 is ..... 45 hours

From these it appears that time taken for getting the usual weighted estimates is about 5/4 of the time taken to get systematic pps estimates. This gap will be large when one has to get estimates for a number of characteristics. In such a case getting the weighted estimates may take about four times the time taken for getting systematic pps estimates. From the point of view of time, there is not much to choose between pps systematic and pps with replacement.

REFERENCES

1. CHURCH, B. N. (1954) : Problems of sample allocation and estimation in an agricultural survey, Journal. Roy. Stat. Soc., 16B, 223-235.

2. HANSEN, M. H., HURWITZ, W.N. and MADOW, W.G.(1955) : Sample Survey Methods and Theory, John Wiley and Sons.

3. LAHIRI, D. B. (1954) : Technical paper on some aspects of development of sample design, Sankhya, 14, 264-316.

4. MURTHY, M. N. and SETHI, V. K. (1959) : Self-weighting design at tabulation stage, National Sample Survey Working Paper No. 6.

5. MURTHY, M. N. and SETHI, V.K. (1961) : Randomized rounded-off multipliers in sampling theory, accepted for publication Jour. Amer. Stat. Assn., 56.

6. SOM, R. K. (1959) : Self-weighting design with an equal number of ultimate stage units in each of the selected penultimate stages units, Bulletin of Calcutta Statistical Association, 9, 59-66.

7. TUKEY, J. W. (1948) : Approximate weights, Ann. Math. Stat., 19, 91-92.

Chapter 8

THEORY OF SURVEY ERRORS

1. INTRODUCTION

In this chapter a comprehensive treatment of the theory of survey errors including both sampling and non-sampling errors is given by bringing together the work of a number of authors on this subject. The use of 'conditional approach' has considerably simplified the derivation of a number of results in this field. Various aspects of the probalem of non-sampling errors such as sources of non-sampling errors, non-sampling bias and variation and non-response are given here. The technique of interpenetrating sub-samples and its various uses in large scale surveys have been discussed in section 11 of this chapter.

The question of assessment and control of non-sampling errors has been receiving considerable attention and suitable techniques are being developed for this purpose. Mahalanobis (1940, 1944, 1946), Mahalanobis and Mahiri (1960) and Lahiri (1957) have given many important techniques for assessing and controlling errors in censuses and surveys. Hansen and others (1946, 1951, 1960) and Sukhatme and Seth (1952) have considered in detail the question of non-sampling errors and have developed a suitable mathematical model for it.

Post-enumeration checks and re-in erview surveys are being made part of some of the nation-wide censuses and surveys so as to enable assessment of non-sampling errors.

2. SOURCES OF NONSAMPLING ERRORS

Till recently the theory of sampling has been developed assuming that each unit in the population has a unique 'true' value and that it can be observed and tabulated without introducing any error. This would mean that a complete enumeration of all the units in the population would give us figures without any errors, which is not usually the case. The above assumption is rather unrealistic, as in practice there are bound to be some observational and tabulation errors in the final results. Of course in some cases these errors may be negligible in the context of the use to which the results are to be put. Even the first part of the assumption that each unit has a unique true value is questionable. As these types of errors are different from the error due to sampling of units and are due to sources other than sampling units, they are termed 'non-sampling errors' or 'response errors'.

The broad sources of non-sampling errors, which are present in both complete enumeration and sample survey, though possibly to varying degrees, are incomplete coverage of the population or sample (including non-response), faulty definitions, defective methods of

data collection and tabulation errors. In case of sample surveys, the errors may also arise from defective sampling frame and selection procedures. More specifically the non-sampling errors may arise due to omission or deplication of units, inaccurate and inappropriate methods of measurement, inappropriate arrangement or wording of questions, inadequate and ambiguous instructions, non-response, deliberate or unconscious misreporting of data by respondents, carelessness on the part of the investigators and clerks, lack of proper supervision, and defective methods of scrutiny and tabulation of data.

From what has been stated above, it is clear that the results of sample surveys are subject not only to sampling error but to non-sampling errors also. In many situations the non-sampling errors may even be larger and therefore more important than the sampling error. Though data obtained on the basis of a complete enumeration are free from sampling error, they are subject to non-sampling errors. To make the results of censuses and surveys useful, it is necessary to reduce the non-sampling as much as possible. It may be noted that while, in general, sampling error decreases with increase in sample size, the non-sampling errors tend to increase with the sample size.

3. CONCEPTUAL SET-UP.

The difference between the sample survey estimate and the parameter true value being estimated may be termed 'error'. If the units in the sample can be observed and tabulated accurately then this error consists of only the error due to sampling, namely,

'sampling error'. A measure of the sampling error is supplied by the mean square error which is the expected value of the square of the difference between the estimator and the true value. This mean square error is composed of two parts - 'sampling bias' and 'sampling variance'. The former has been defined as the difference between the expected value of the estimator and the true value and the latter is a measure of the divergence of the estimator from its expected value. Of course in some cases the sampling bias may be negligible or zero.

If the data are also subject to non-sampling errors, then the difference between a survey estimate and the parameter true value may be termed 'total error' and this consists of both sampling and non-sampling errors. In this section a conceptual set-up is developed which would enable us to get a measure of the non-sampling errors in terms of 'non-sampling bias' and 'non-sampling variance'.

The 'true' value of a unit is to be conceived of as a characteristic of the unit independent of the survey conditions which may affect the value 'reported' for that unit. For instance age of a person at a particular point of time, income of a person during a particular period of time and number of persons in a country at a point of time are examples of characteristics for which the true value exists and is clearly defined. There are many items of information, such as intelligence of a person, attitude to some social

measures, consumer preference to certain articles, for which it is very difficult even to conceive of the true value. In such cases some suitable conceptually and to some extent arbitrarily defined value may be taken as the true value. For the definition of a true value to be useful in practice, it should serve the purpose of the survey and it should be well defined and observable under 'reasonable conditions of survey' relating to subject coverage, method of enquiry survey period and method of tabulation. The non-sampling errors arise due to the fact that it may not be possible to collect and process data accurately even if the true value is well defined because of mm so many operational difficulties.

Suppose a sample has been chosen to be canvassed under reasonable conditions of survey and that there are two populations, one of investigators and the other of tabulators (clerks) qualified for doing the field and processing work of the survey. If we repeatedly carry out the survey on the selected units with different samples of investigators and computors chosen with some suitable probability designs, we may get different results because of the various possible sources of error present under the usual operational conditions. Here there are three stages of randomisation - selection of units, investigators and computors. The difference between the expected value of the estimator taken over all the three stages of randomization and the true value may be termed 'total bias'. This consists

of both 'sampling bias' and 'non-sampling bias'. The variance of the estimator taken over all the three stages of randomization measures the divergence of the estimator from its expected value and consists of sampling variance, variance between investigators, variance between computors and some interactions between the three sources of error. For instance the data collected by one investigator may be affected by his misunderstanding of the instructions, his preconceived notions about the survey, the earlier units canvassed by him etc. Thus we see that the total error consists of sampling bias and variance, non-sampling or response bias and variance and some interactions between the sample and the sources of non-sampling errors.

4. NON-SAMPLING BIAS

For the sake of simplicity let us assume only two stages of randomization - one for selecting the sample of units and the other for selecting the survey personnel - instead of the three stages of randomization considered earlier. Here we consider the survey personnel as a whole instead of as investigators and computors. Let $\hat{Y}_{sr}$ be the estimate for the sth sample of units supplied by the rth sample of the survey personnel. The conditional expected value of $\hat{Y}_{sr}$ taken over the second stage of randomization for a fixed sample of units is given by

$$E_r(\hat{Y}_{sr}) = \hat{Y}_s. \qquad (4.1)$$

which may be different from the estimate $\hat{Y}_s$ based on the true values of the units in the sample. The expected value of this $\hat{Y}_s$ over the first stage of randomization gives

$$E_s(Y_s) = Y' \qquad (4.2)$$

which is the value π that can be unbiasedly estimated by the specified survey process. This value Y' may be different from the true population total Y and the difference

$$B(t) = Y' - Y \qquad (4.3)$$

may be termed the 'total bias'.

It may be noted the sampling bias is given by

$$B(s) = E_s(Y_s) - Y \qquad (4.4)$$

which is the difference between the expected value of the estimator based on the true values and the true value of the population total. Since the total bias is the sum of sampling and non-sampling biases, the non|sampling or response bias is given by

$$B(r) = B(t) - B(s) = Y' - E_s(\hat{Y}_s) = E_s(\hat{Y}_{s.} - \hat{Y}_s) \qquad \ldots (4.5)$$

which is the expected value of the non-sampling or response deviation for the sth sample. If it is a complete enumeration, there is no sampling bias and the total bias consists of only the response bias.

In case of many sample surveys also, the total bias consists of only the response bias, since usually unbiased estimators (from the point of view of sampling units) are used.

To fix the ideas let us consider an example where a simple random sample of units is drawn with or without replacement from a population of N units and surveyed by k persons chosen with equal probability from a large population of K persons qualified for this work, each person surveying m of the units assigned to him at random (n=mk). Let y_{sij} be the value reported by the jth investigator for the ith unit allotted to him in the sth sample. Then an estimator of the population total is given by

$$Y = \frac{N}{n} \sum_{j}^{k} \sum_{i}^{m} y_{sij} \qquad (4.6)$$

The conditional expected value over all possible samples of investigators where the sth sample is fixed is

$$E(\hat{Y}/s) = \frac{N}{n} \sum_{i}^{n} y_{si.} \quad (n = mk), \; y_{si.} = \frac{1}{K} \sum_{j}^{K} y_{sij}$$

If there were no non-sampling errors in the survey, the estimator would be

$$Y = \frac{N}{n} \sum_{i}^{n} y_i$$

where y_i is the true value. The difference

$$d_{si} = y_{si.} - y_i \qquad (4.7)$$

may be considered to be the response deviation. This deviation may also depend on the particular sample being surveyed because of the

possible influence of some units on those of the others in the sample. The response bias in this case is given by

$$B(r) = Y' - Y \tag{4.8}$$

where

$$Y' = \frac{1}{\binom{N}{n}} \frac{N}{n} \sum_{i} \sum_{s \ni i} Y_{si}.$$

$\sum_{s \ni i}$ standing for the summation over all samples containing the ith unit. If the average response for a unit is not affected by that of another unit in the sample, i.e. if $y_{si} = y'_i$, then

$$Y' = \frac{1}{N} \sum_{i}^{N} Y'_i, \quad (Y_i = \frac{1}{K} \sum_{j}^{K} Y_{ij}) \tag{4.9}$$

There are a number of techniques available for the assessment of response bias (Lahiri, 1957). The survey figure may be compared with an external figure obtained by some other agency or by the same agency in some previous period after making the necessary adjustments for differences in coverage, definitions, survey period etc.

This comparison may be taken as a broad check on the survey figure. This check is termed 'external aggregative check'. A better check would be to have unit by unit comparison of the survey data with the corresponding values in some other survey. This method is termed 'external unitary check'. It may be noted that there would be considerable difficulties in matching the units for this type of

check. In these checks the assumption is that one source of data is more reliable than the other. If this assumption is not true, it would be difficult to conclude which figure is subject to more bias in case of discrepancies. Another technique of assessing response bias is to draw the sample in the form of two or more interpenetrating sub-samples and to get these surveyed by different groups of investigators. This procedure is known as the method of interpenetrating sub-samples and will be considered in detail in section 11.

The response bias in a census can be estimated by surveying a sample of units in the population using better techniques of data collection and compilation than would be possible under census conditions. Such surveys which are usually conducted just after the census to study the quality of the census data are called 'post enumeration surveys'. Even in case of a large scale sample survey, the response bias can be estimated by resurveying a sub-sample of the original sample using better survey techniques. Another method of checking survey data would be to compare the values of the units obtained in two surveys and to reconcile the figures by further investigation in case of discrepancies. This method of checking is termed 'reconciliation (check) surveys'.

5. NON-SAMPLING VARIANCE.

The mean square error of the estimator Y_{sr} based on the sth sample of units and supplied by the rth sample of the survey personnel, is by definition

$$M(\hat{Y}_{sr}) = E_{sr}(Y_{sr} - Y)^2 \qquad (5.1)$$

where Y is the true value being estimated. This is a measure of the divergence of the estimator from the true value, taking into account both sampling and non-sampling errors. This measure consists of bias and variance, that is,

$$M(Y_{sr}) = V(Y_{sr}) + B^2(Y_{sr})$$

$$= E(Y_{sr} - Y')^2 + (Y' - Y)^2 \qquad (5.2)$$

where Y' is the expected value of the estimator taken over both the stages of randomization. The variance of the estimator is a measure of the divergence of the estimator from its expected value and $Y' - Y$ is the bias. Taking the variance over the two stages of randomization, we get

$$V_{sr}(Y_{sr}) = V_s E_r(Y_{sr}) + E_s V_r(Y_{sr})$$

$$= V_s(Y_s) + E_s E_r(Y_{sr} - Y_s)^2 \qquad (5.3)$$

From (5.3) we see that the variance can be split up into two parts- sampling variance and response variance. The second term stands for the expected value of the square of the response deviations of the sample estimates from their expected value taken over both the stages of randomization. This term can be further split up by writing

$$Y_{sr} - Y_{s.} = (Y_{sr} - Y_{s.} - Y_{.r} + Y') + (Y_{.r} - Y')$$

where $Y_{.r} = E_s(Y_{sr})$, and taking the variance we get

$$E_{sr}(Y_{sr} - Y_{s.})^2 = E_{sr}(Y_{sr} - Y_{s.} - Y_{.r} + Y')^2 + E_r(Y_{.r} - Y')^2 \quad \ldots (5.4)$$

The first term on the right in (5.4) is the interaction between the sampling and non-sampling errors and the second term is the variance between survey personnel. Thus we see that the mean square of the estimator consists of sampling variance, interaction between sampling and non-sampling errors, variance between survey personnel and square of sum of the sampling and non-sampling biases. In a complete census the mean square error is composed of only the non-sampling variance and square of the response bias.

6. SIMPLE RANDOM SAMPLING

To fix the ideas let us consider the case where a simple random sample of n units drawn with replacement from a population of N units is divided at random into k equal sub-samples of m units each and these sub-samples are surveyed by k investigators selected with equal probability from a large population of K investigators qualified for this work. Let Y_{ij} and Y_i be the value reported by the jth investigator for the ith unit in the population and its true value respectively. Suppose Y_{ij} is the value reported for the ith unit in the sample by the jth selected investigator. Here it is assumed that the response for a unit is not affected by the response of other units in the sample. An estimator of the popula-

tion mean is given by

$$\bar{y} = \frac{1}{n} \sum_{j}^{k} \sum_{i}^{m} y_{ij} \text{ , } (n = km). \qquad (6.1)$$

The expected value of the estimator taken over the two stages of randomization is

$$E(\bar{y}) = \frac{1}{N} \sum_{i}^{N} Y_i' \text{ , } (Y_i' = \frac{1}{K} \sum_{j}^{K} Y_{ij}) \qquad (6.2)$$

and the total bias, which in this case consists wholly of response bias, is

$$B(t) = B(r) = \frac{1}{N} \sum_{i}^{N} (Y_i' - Y_i) \qquad (6.3)$$

The variance of the estimator is given by

$$V_{sr}(\bar{y}) = V_s E_r(\bar{y}) + E_s V_r(\bar{y})$$

where the subscripts denote the stages of randomization. The conditional expected value of y over the second stages of randomization for a fixed sample of units is

$$E_r(\bar{y}) = \frac{1}{n} \sum_{i}^{n} y_i' \text{ } (y_i' = \frac{1}{K} \sum_{j}^{K} y_{ij}).$$

The unconditional variance of this over the first stage of randomization is given by

$$V_s E_r(\bar{y}) = \frac{\sigma_s^2}{n} \text{ , } \sigma_s^2 = \frac{1}{N} \sum_{i}^{N} (Y_i - Y')^2 \qquad (6.4)$$

The conditional variance of y over the second stage of randomization for a fixed sample of units is

$$V_r(y) = \frac{1}{k} E_r\left(\frac{1}{m}\sum_{i}^{m} y_{ij} - \frac{1}{m}\sum_{i}^{m} y'_i\right)^2$$

$$= \frac{1}{km^2}\,\frac{1}{K}\sum_{j}^{K}\left[\sum_{i}^{m}(y_{ij}-y'_i)^2 + \sum_{i}^{m}\sum_{i'\neq i}(y_{ij}-y'_i)(y_{i'j}-y'_i)\right]$$

for $V_r\left(\frac{1}{m}\sum_{i}^{m} y_{ij}\right)$ is the same for all j. Taking the unconditional expected value of this expression over the first stage of randomization, we get

$$E_s V_r(\bar{y}) = \frac{1}{km^2}\,\frac{1}{K}\sum_{j}^{K}\left[\frac{m}{N}\sum_{i}^{N}(Y_{ij}-Y'_i)^2 + \frac{m(m-1)}{N(n-1)}\sum_{i}^{N}\sum_{i'\neq i}\right.$$

$$\left.(Y_{ij}-Y'_i)(Y_{i'j}-Y'_{i'})\right]$$

$$= \frac{1}{km}\,\sigma_d^2\,[1+(m-1)\,\rho] \qquad (6.5)$$

where σ_d^2 is termed 'simple' or 'uncorrelated' response variance and is given by the variance of individual response deviations, that is,

$$\sigma_d^2 = \frac{1}{KN}\sum_{i}^{N}\sum_{j}^{K}(Y_{ij}-Y'_i)^2 \qquad (6.6)$$

and ρ is the intra-class correlation among the response deviations in a sample canvassed by one investigator (intra-investigator correlation), and is given by

$$\rho = \frac{1}{KN(N-1)}\sum_{j}^{K}\sum_{i}^{N}\sum_{i'\neq i}(Y_{ij}-Y'_i)(Y_{i'j}-Y'_{i'}) \qquad \ldots \qquad (6.7)$$

Hence the variance and the mean square error of $\bar{y}$ are

$$V(\bar{y}) = \frac{\sigma_s^2}{n} + \frac{\sigma_d^2}{n}\left[1 + (m-1)\rho\right] \qquad (6.8)$$

and

$$M.S.E.(\bar{y}) = V(\bar{y}) + (\bar{Y}' - \bar{Y})^2. \qquad (6.9)$$

In case of complete census, sampling variance would be zero and hence the variance and mean square error of the census figure is given by

$$V(\bar{y}) = \frac{\sigma_d^2}{N}\left[1 + (m-1)\rho\right] \qquad (6.10)$$

$$M.S.E.(\bar{y}') = V(\bar{y}') + (\bar{Y}' - \bar{Y})^2 \qquad (6.11)$$

The result in (6.8) shows the contribution to the total variance from the response variation and it also brings out the impact of the intra-class correlation among the responses in a sample canvassed by one investigator (intra-investigator correlation) on the response variance. The intra-class correlation will be positive if the response deviations for the different units have a consistent tendency to be in one direction for an investigator and in another direction for another investigator. Even when this correlation is small, the contribution to the response variation may be considerable if m, the number of units surveyed by each investigator is large. For instance, if $\rho = 0.01$ and $m = 1000$, the response variation becomes about ten times more than that in case of $\rho = 0$.

An unbiased estimator of the variance of the estimator $\bar{y}$ given in (6.8) is given by

$$\hat{V}(\bar{y}) = \frac{1}{k(k-1)} \sum_{j}^{k} (\bar{y}_{.j} - \bar{y})^2, \quad (\bar{y}_{.j} = \frac{1}{m} \sum_{i}^{m} y_{ij}) \qquad (6.12)$$

for

$$E\left(\sum_{j}^{k} (\bar{y}^2_{.j} - k\bar{y}^2) \right) = k \left[k\, V(\bar{y}) + \bar{y}^2 - V(\bar{y}) - \bar{y}^2 \right] = k(k-1)V(\bar{y}).$$

This MMM result shows that if k independent samples are surveyed by k investigators selected with equal probability from a large population of investigators, then it is possible to get an unbiased estimator of the total variance (and not the total mean square error). This procedure is known as the method of 'interpenetrating sub-samples' which is considered in detail in section 10. The variance between investigators is given by

$$\sigma_r^2 = \frac{1}{K} \sum_{j}^{K} (\bar{Y}'_{.j} - \bar{Y})^2 - \sigma_d^2 \qquad (6.13)$$

for

$$\sigma_r^2 = \frac{1}{K} \sum_{j}^{K} \left[\frac{1}{N} \sum_{i}^{N} (Y_{ij} - Y'_i) \right]^2 = \frac{1}{KN^2} \sum_{j}^{K} \sum_{i}^{N} (Y_{ij} - Y'_i)^2 +$$

$$\frac{1}{KN^2} \sum_{j}^{K} \sum_{i}^{N} \sum_{i' \neq i} (Y_{ij} - Y'_i)(Y_{i'j} - Y'_{i'})$$

$$= \frac{\sigma_d^2}{N} + \frac{N-1}{N} \sigma_d^2 = \sigma_d^2$$

if N is assumed to be large. An unbiased estimator of σ_r^2 is given by

$$\hat{\sigma}_r^2 = k\, \hat{V}(\bar{y}) - \frac{1}{km(m-1)} \sum_{j}^{k} \sum_{i}^{m} (y_{ij} - \bar{y}_{.j})^2 \qquad (6.14)$$

for taking the conditional expected value of the second term in (6.14) we get

$$\frac{1}{nk} \sum_{j}^{k} \frac{1}{N} \sum_{i}^{N} (Y_{ij} - Y'_{.j})^2$$

and the expected value of this expression over the sample of investigators, is given by

$$\frac{1}{n} \frac{1}{NK} \sum_{j}^{K} \sum_{i}^{N} (Y_{ij} - \bar{Y}'_{.j})^2 = \frac{1}{n} (\sigma^2 - \sigma_r^2)$$

where σ^2 is the total variance in the population and is given by

$$\sigma^2 = \frac{1}{KN} \sum_{j}^{K} \sum_{i}^{N} (Y_{ij} - Y')^2 = \sigma_n^2 + \sigma_d^2 \qquad (6.15)$$

Hence

$$E(\sigma_r^2) = k \left[\frac{\sigma^2}{nk} + \frac{(n-1)}{nk} \sigma_r^2 \right] - \frac{1}{n} (\sigma^2 - \sigma_r^2) = \sigma_r^2 .$$

7. ESTIMATION OF POPULATION PROPORTION.

It is interesting to consider the question of response variance in estimating a population proportion. Let Y_{ij} be 1 or 0 according as the jth investigator reports the ith unit in the population as belonging to a particular class or not and let P'_i be the proportion of the investigators reporting the ith unit in the population as belonging to that class. Suppose a simple random sample of units is drawn with replacement from a population of N units to be surveyed by a sample of k persons selected with equal probability

from a large population of K persons qualified for this work. An estimator of the population proportion P is given by

$$\hat{P} = \frac{1}{mk} \sum_{j}^{k} \sum_{i}^{m} y_{ij} \quad (n = mk) \tag{7.1}$$

where y_{ij} is 1 or 0 according as the jth selected investigator reports the ith unit in the sample as belonging to a given class or not. The expected value of this estimator over both the stages of randomisation is

$$E(\hat{P}) = \frac{1}{N} \sum_{i}^{N} P'_i = P' \tag{7.2}$$

and the bias, which in this consists of only the response bias, is $P' - P$. In this case σ_s^2 and σ_d^2 defined in (6.4) and (6.6) respectively are given by

$$\sigma_s^2 = \frac{1}{N} \sum_{i}^{N} (P'_i - P')^2 \tag{7.3}$$

and

$$\sigma_d^2 = \frac{1}{N} \sum_{i}^{N} P'_i Q'_i, \quad (Q'_i = 1 - P'_i) \tag{7.4}$$

The variance of P is

$$V(P) = \frac{1}{Nn} \sum_{i}^{N} (P'_i - P')^2 + \frac{1}{Nn} \sum_{i}^{N} P' Q' [1 + (n-1)\rho] \tag{7.5}$$

From (6.12) it can be seen that an unbiased estimator of the total variance given in (7.5) is

$$\hat{V}(\hat{P}) = \frac{1}{k(k-1)} \sum_{j}^{k} (p_{.j} - p)^2 \tag{7.7}$$

where $p_{.j}$ is the sample proportion reported by the jth selected investigator in the sample assigned to him and p is the over-all sample proportion. From (5.14) an unbiased estimator of the variance between investigators σ_r^2 is given by

$$\sigma_r^2 = V(P) - \frac{1}{m(m-1)k} \sum_j^k p_{.j}\, q_j \,(q_{.j} = 1 - p_{.j}) \qquad (7.7)$$

If the intra-class correlation is assumed to be 0, then the variance given in (7.5) reduces to

$$V(P) = \frac{P'Q'}{n} \quad (Q' = 1 - P') \qquad (7.8)$$

This result is interesting because it shows that the expression which is normally used as the sampling variance of a sample proportion includes not only the sampling variance but also the uncorrelated response variance (Hansen, Hurwitz and Bershad, 1960). An unbiased estimator of the variance is

$$V(P) = \frac{pq}{(n-1)}, \quad (q = 1-p)$$

since $E(pq) = E(p) - E(p^2) = P' - V(p) - p'^2 = (n-1)V(p)$. Here again we see that the variance estimator of a sample proportion usually used to estimate the sampling variance estimates unbiasedly the total variance including both the sampling variance and the uncorrelated response variance.

8. COST FUNCTION

Let us consider the case of getting optimum values of k, the number of investigators, and m, the number of units assigned to one

investigator, which would minimize the total variance for a given fixed cost. Suppose the cost function

$$C = kC_1 + nC_2 \tag{8.1}$$

where C_1 is the cost of recruiting and training one investigator, C_2 is the cost of surveying one unit and $n = km$. The total variance of the estimator y of the population mean Y, given in (6.8), may be written as

$$V(y) = \frac{\sigma^2 - \sigma_r^2}{n} + \frac{\sigma_r^2}{k} \tag{8.2}$$

where $\sigma_r^2 = \sigma_d^2$ and $\sigma^2 = \sigma_s^2 + \sigma_d^2$. Minimising the variance in (7.2) with respect to n and k subject to the cost restriction (7.1), we get

(8.3)

$$m = \frac{C_1}{C_2} \quad \frac{\sigma^2 - \sigma_r^2}{\sigma_r^2} \tag{8.4}$$

9. INTRA-INVESTIGATOR CORRELATION.

A number of empirical studies have been conducted in recent years to assess the magnitude of the intra-investigator correlation coefficient for different types of characteristics. The results obtained in some of these studies are presented in Table 1.

Table 1. Showing the ranges of the intra-investigator correlation coefficient for different types of characteristics.

reference	k	type of items	range of
(1)	(2)	(3)	(4)
Gray (1956)	20	8 factual items	0 - 0.02
		perception of and attitude to neighbour's noises	0 - 0.08
		8 items about illness	0 - 0.11
Gales and Kendall (1957)	48	most semi-factual and attitudinal items about TV habits	0 - 0.05
Hansen and Marks (1958)	705	most factual items	0.01 - 0.02
		more difficult items	0.02 - 0.04
		most 'not ascertained' categories	0.02 - 0.06
Kish and Slater (1960)	20	first study	0 - 0.07
	8	second study	0 - 0.04
Kish and Rangford (in preparation)	9	measurement of major dental defects	0.15 - 0.50

number of investigators = k.
(Source : Kish, L and Slater, C.W.(1960) 'Two studies of interviewer variance of socio-psychological variables', presented at the Annual Meeting of the American Statistical Association).

10. NON-RESPONSE ERROR.

One of the sources of error in censuses and surveys, mentioned earlier, is the incomplete coverage of the population or sample. This

incomplete coverage may arise due to respondents' refusal to give information, respondents being 'not at home', inaccessible sample units etc. The error in this case arises because the population of non-respondents may have characteristics different from those who respond and the results based only on the surveyed units may be misleading. This type of error may be termed 'non-response error', since it arises from not surveying all the units in the population of sample. The non-response error may not be important if the units not responding in a survey have characteristics similar to those of the responding units. But usually in practice this situation does not arise. For instance, if questionnaires are mailed to a number of farmers, the non-response rate may not be uniform among the farmers belonging to different land holding size classes and hence the results based only on the responses of the responding farmers may be misleading. It may be noted that in most cases of non-response, the response may be obtained by pursuation, repeated visits to the non-responding units etc.

One way of dealing with the problem of non-response is to make all efforts to collect information from a sub-sample of the units not responding in the first attempt (Hansen and Hurwitz, 1946). Suppose out of n units selected with equal probability without replacement from a population of N units, n_1 units respond and n_2 $(=n-n_1)$ units do not respond in the first attempt. Let a sub-sample

of n'_2 units be selected from the n_2 non-responding units with equal probability without replacement for making special efforts to collect the information. If y_1 and y'_2 are sample means based on the n_1 units responding in the first attempt and on the sub-sample of n'_2 units respectively, then an unbiased estimator of the population total Y is given by

$$Y = \frac{N}{n}(n_1 y_1 + n_2 y'_2) \quad . \tag{10.1}$$

It may be noted that there are three stages of randomization in this case - sampling of units, number of units in the sample not responding in the first attempt and sub-sampling of n'_2 units from the n_2 units not responding in the first attempt. Taking the variance of the estimator given in (10.1) over these three stages of randomization, we have

$$V_{123}(\hat{Y}) = V_1 E_2 E_3(\hat{Y}) + E_1 V_2 E_3(\hat{Y}) + E_1 E_2 V_3(\hat{Y})$$

where E and V stand for conditional expected value and variance and the subscripts denote the stages of randomization. The conditional expected value and variance over the third stage of randomization are given by n_2

$$E_3(\hat{Y}) = \frac{N}{n}(n_1 y_1 + n_2 y_2) = \frac{N}{n}, \quad (y_2 = \frac{1}{n_2} \sum_i y_{2i})$$

and

$$V_3(\hat{Y}) = \frac{N^2}{n^2} n_2(n_2 - n'_2) \frac{s_2^2}{n'_2}, \quad s_2^2 = \frac{1}{(n_2-1)} \sum_i^{n_2} (y_{2i} - \bar{y}_2)^2$$

where $\bar{y}$ is the sample mean based on all the n units in the sample and y_{2i} is the value of the ith non-responding unit in the sample. Further it can be seen that

$$E_2\left(\frac{N}{n}\bar{y}\right) = N\bar{y} \text{ and } V_2(N\bar{y}) = 0.$$

Hence the variance of the estimator is given by

$$V(\hat{Y}) = V_1(N\bar{y}) + E_1E_2\left[\frac{N^2}{n}\left(\frac{n_2}{n}\right)(k-1)\,s_2^2\right]$$

where $k = n_2 / n'_2$. Since

$$E_1(s_2^2) = \frac{N_2}{N_2-1}\sigma_b^2 \quad \text{and} \quad E_2\left(\frac{n_2}{n}\right) = \frac{N_2}{N}$$

where σ_b^2 is the variance between the units in the population not responding in the first attempt and N_2 is the number of such non-responding units in the population, the variance becomes

$$V(\bar{Y}) = \frac{N^2(N-n)}{(N-n)}\,\frac{\sigma^2}{n} + \frac{N}{n}(k-1)\,\frac{N_2^2\,\sigma_b^2}{(N_2-1)} \qquad (10.2)$$

where σ^2 is the variance between the N units in the population.

The cost function in this case may be of the form

$$C = C_1n + C_2nP + C_3\frac{n}{k}Q \qquad (10.3)$$

where C_1 is the cost per unit for the first attempt at data collection, C_2 is the cost per unit for tabulation, C_3 is the cost per unit sampled from the non-responding units (for obtaining data by additional efforts and for tabulation), P is the proportion of units in the population that would have responded in the first attempt, and $Q = 1 - P$. The optimum values of n and k which would minimize

the cost, ensuring at the same time a given value v^2 for the variance of the estimator, are given by

$$n = \hat{n}\left[1 + (k-1)\,Q^2\,\frac{N-1}{N_2-1}\,\frac{\sigma_b^2}{\sigma^2}\right] \qquad (10.4)$$

$$k = \left[\frac{N^2(N_2-1)\sigma^2}{N_2^2(N-1)\sigma_b^2} - 1\right]\frac{C_3 Q}{C_1 + C_2 P} \qquad (10.5)$$

where

$$\hat{n} = \frac{N\sigma^2}{\sigma^2 + \frac{N-1}{N^2}v^2}$$

the sample size required for ensuring the value v^2 for the variance if there were complete response. If it is assumed that $\sigma^2 = \sigma_b^2$ and $\frac{N}{N-1} = \frac{N_2}{N_2-1} = 1$, the optimum values of n and k reduce to

$$n = \hat{n}\left[1 + (k-1)Q\right] \qquad (10.6)$$

$$k = \frac{C_3 P}{C_1 + C_2 P}.$$

An interesting device in dealing with 'not at home' cases has been considered by Politz and Simmons (1949). This procedure consists in ascertaining from the responding households the chance of their being at home at a particular point of time and weighting the results with the inverse of this chance. For instance the households may be asked whether they were at home at some specified time during the previous 7 days. Then the households may be classified as being at home once in 6 visits, twice in 6 visits etc), and the

data obtained for the different classes may be weighted by the inverse of the respective probabilities of being at home. In practice some bias would still persist because of persons not at home during the entire investigation period, who cannot be contacted.

11. INTERPENETRATING SUB-SAMPLES

The technique of interpenetrating sub-samples, which is due to Mahalanobis (1944, 1945, 1946), in its most general sense consists of drawing the sample in the form of k sub-samples according to any probability sampling design which would enable in getting valid estimates of the population parameter under consideration and subjecting these to k different operations to study the differential effects of these operations. This technique has many possibilities in the field of censuses and surveys in assessing non-sampling errors (Mahalanobis and Lahiri, 1960, Lahiri, 1953, 1957). One of the advantages has been mentioned in section 6. There it was shown that if k independent interpenetrating sub-samples drawn from a population are assigned at random to k investigators selected with equal probability from a large population of investigators, it would be possible to estimate the total variance of the estimator including both sampling and response variation.

11.1 LINKED SUB-SAMPLES

Originally Mahalnobis (1940) made use of this technique in crop surveys to find out the differential investigator bias. For this purpose, linked pairs of grids (square parcel of land) were lo-

located at random on the maps in the form of dumb-bell shaped figures, one end of each figure representing the grid belonging to sub-sample 1 and the other end representing the grid belonging to sub-sample 2. One sub-sample was investigated by one set of investigators and the other sub-sample by an entirely different set of investigators independently. Under certain well-known assumptions the Student's t-test may be applied to the difference between the estimates based on the two sub-samples to test the hypothesis that there is no differential investigator bias at any specified level of significance. If the difference turns out to be significant, it means that the direction and magnitude of investigator bias are not of the same order for all the investigators. It may be noted that if the difference turns out to be statistically insignificant, it does not mean that the investigator bias is zero. For, this result may be due to the fact that the investigator biases are all of the same order and in the same direction.

The above method can well be applied to bring out the differential effect of different tabulation procedures, methods of data collection, etc., and to bring out the variation over time. Suppose one is interested in finding out whether intensive training of the investigators for a given survey is essential or not. For this purpose, one sub-sample may be assigned to intensively trained investigators and the other sub-sample to investigators who have got

only superficial training. If the difference in the results obtained from these two sub-samples turns out to be significant, there is a strong case for adopting the method of intensive training in future surveys of a similar nature. On the other hand if the difference were not significant, it would mean that for this type survey intensive training is perhaps not essential.

The technique of interpenetrating sub-samples may be used as a check on the different operations involved in large scale surveys. Suppose one wishes to have a check on the calculations at the time of tabulation. For this purpose, the sample may be divided into k linked sub-samples assigned to k different groups of computors at random and the estimates may be obtained from each of these sub-samples. If there is good agreement between these estimates, for all practical purposes it may be assumed with certain amount of confidence that the calculations have been done correctly. If one of these estimates differs from the others (assuming k is more than 2) and if there is good agreement between the remaining k-1 estimates, one naturally suspects the calculations done on that sub-sample and gets that estimate recalculated. Thus it is seen that suitable action can be taken on the basis of the sub-sample estimates thereby increasing the accuracy and utility of the final results.

It is to be noted that detailed interpenetration of the sub-sample would require a good deal of additional preparatory time, and

increases the complexity of work at the field and the tabulation stages. It is also found that the power of the interpenetrating sub-sample check is generally low due to the fact that the estimate of variance usually used in the test is based only on a few degrees of freedom. It may be noted that the larger the positive correlation between the sub-samples, the greater will be the sensitivity of the test and lower will be the efficiency of the joint estimate based on both the sub-samples. So if the main object of the survey is to test the differential investigator bias, then it would be desirable to have the same sample investigated by both the sets of investigators independently under the same conditions. If this is not possible due to the presence of conditioning effect between successive investigations of the same unit, it would be desirable to use sub-samples which are linked in such a way that the estimates from these samples are highly correlated.

Another objection to the use of interpenetrating sub-samples is that the cost of survey increases because of both the parties of investigators going over the entire field. It has been shown that under the assumption that the cost of journey is proportional to the square-root of the number of randomly located points the percentage loss of information (L_p) per unit of cost resulting from the linked method of selection as compared to selection without interpenetration for the same total sample size is given by

$$L_p = \left[1 - \frac{(C_j + C_k)}{(1+\rho)(2C_j + C_k)} \right] \times 100 \qquad (11.1)$$

where C_j is the cost of journey, C_k other costs, ρ the correlation coefficient between the two linked samples. Information is defined as a quantity which varies inversely as the variance of the estimate and the ratio of the information to the total cost of the survey is defined as the 'information per unit cost' (Mokashi, 1950).

11.2 INDEPENDENT SUB-SAMPLES.

As has already been pointed out, linked samples are to be used only if the main objective is to find out the differential effect of two operations. But if the main object is to get a reliable estimate of the population parameter and the study of differential effects is only a subsidiary objective, then it is preferable to have independent interpenetrating sub-samples. The difference between the estimates based on two independent interpenetrating sub-samples provide a measure of the sampling as well as non-sampling errors present in the results.

The technique of interpenetrating sub-samples is of held in calculating the total variation especially in large scale sample surveys where a number of characteristics are under consideration. If there are k independent interpenetrating sub-samples subjected to k different operations each providing a valid estimate of the population

parameter under consideration, then an unbiased estimator of the variance of the estimator (mean of the sub-sample estimates) is given by

$$v(\bar{y}) = \frac{1}{k(k-1)} \sum_{i}^{k} (y_i - \bar{y})^2, \quad (\bar{y} = \frac{1}{k} \sum_{i}^{k} y_i) \qquad (11.2)$$

where y_i is the estimate based on the ith sub-sample. It may be noted that this procedure gives a simple method of getting an estimator of the variance of a ratio estimator. If r_i $(= \frac{y_i}{x_i})$, $(i = 1, 2, \ldots, k)$ is an estimate of the population ratio $R(= \frac{Y}{X})$ based on the ith sub-sample, then an unbiased estimator of the variance of

$$R' = \frac{1}{k} \sum_{i}^{k} r_i \qquad (11.3)$$

is given by

$$\hat{v}(R') = \frac{1}{k(k-1)} \sum_{i}^{k} (r_i - R')^2 \qquad (11.4)$$

Since the variance of R'_k and that of the combined ratio estimator

$$R'' = \frac{\sum_{i} y_i}{\sum_{i} y_i} \qquad (11.5)$$

are approximately the same (Murthy and Nanjamma, 1959), (10.4) can be taken as an estimator of the variance of R''.

It is to be noted that the variance estimator given in (11.2) holds even if the variances of sub-sample estimates are different provided ꝛ the combined estimator is taken as the arithmetic mean

of the sub-sample estimates. An unbiased estimator of variance can be obtained on the basis of independent interpenetrating sub-sample estimates even if the sub-sample estimates are weighted to obtain the combined estimator. If y_i and w_i are respectively the estimates based on and the weight for the ith sub-sample (i=1,2,..., k) then an unbiased estimator of the variance of the combined estimator

$$\hat{Y} = \sum_{i}^{k} w_i y_i, \quad (\sum_{i}^{k} w_i = 1) \tag{11.6}$$

is given by

$$\hat{V}(\hat{Y}) = \hat{Y}^2 - \frac{2}{1 - \sum w_i^2} \sum_{i}^{k} \sum_{j>i} w_i w_j y_i y_j$$

since $E(\hat{Y}^2) = \hat{V}(\hat{Y}) + Y^2$ and the second term in the above expression estimates unbiasedly Y^2. This expression after simplification becomes

$$\hat{V}(\hat{Y}) = \frac{(\sum w_i^2 y_i^2 - (\sum w_i^2)(\sum w_i y_i)^2}{(1 - \sum w_i^2)} \tag{11.7}$$

Since it may be difficult to compute the estimator in practice the following unbiased estimator is suggested.

$$\hat{V}(\hat{Y}) = \bar{y}^2 \tag{11.8}$$

In case of k = 2, this becomes simply

$$V'(\bar{y}) = \bar{y}^2 - y_1 y_2 ,$$

which is quite simple to calculate since $\bar{y}$, y_1 and y_2 would be readily available.

Suppose in a stratified sample design, there are k independent inerpenetrating sub-samples in each stratum. Let y_{si} denote the estimate of the sth stratum total based on the ith sub-sample (s=1,2,... L, i = 1, 2, ..., k). The variance estimator based on sub-sample estimates may be obtained either using strata sub-sample estimates or just the sub-sample estimates pooled over the strata. That is

$$\hat{V}_1(\hat{Y}) = \frac{1}{k(k-1)} \sum_{s}^{L} \sum_{i}^{k} (y_{si} - \bar{y}_{s.})^2 \qquad (11.9)$$

$$\hat{V}_2(\hat{Y}) = \frac{1}{k(k-1)} \sum_{i}^{k} (y_{.i} - \bar{y})^2 \qquad (11.10)$$

where

$$\bar{y}_s = \frac{1}{k} \sum_{i}^{k} y_{si}, \; y_{.i} = \sum_{s}^{L} y_{si} \text{ and } \bar{y} = \frac{1}{k} \sum_{i}^{k} \bar{y}_{.i}$$

Of these two estimators (11.9) is more efficient than (11.10) (Murthy 1961), though the calculation of the latter will be less time consuming than the former. In a stratified sample design with k independent interpenetrating sub-samples if y_{si} and x_{si} denote the estimates of the sth stratum total for the characteristics y and x respectively based on the ith sub-sample, then an estimator of the variance of the ratio estimator $\hat{T}(= y/s)$ is given by

$$v(R) = \frac{1}{\bar{x}^2} \frac{1}{k(k-1)} \sum_{s}^{L} \Big[\sum_{i}^{k} (y_{si} - \bar{y}_s)^2 - 2R \sum_{i}^{k} (y_{si} - \bar{y}_s)(x_{si} - \bar{x}_s) + R^2 \sum_{i}^{k} (x_{si} - \bar{x}_s)^2 \Big] \qquad (11.11)$$

and an estimator of the bias in R is given by

$$B(\bar{R}) = \frac{n\bar{R} - R'}{(n-1)} \qquad (11.12)$$

where $R' = \frac{1}{n} \sum_{i}^{k} \frac{y_{.i}}{x_{.i}}$ (Murthy and Nanjamma, 1959).

Operationally this method is convenient because it simplifies the computation of variance in case of complicated designs and at the same time helps in having a broad internal check on the results. The efficiency of the variance estimator is, however, impaired due to the reduction in the number of degrees of freedom on which such estimates are based. This also makes the large sample interpretation of the variance estimator inapplicable. However the range of the sub-sample estimates provides a confidence interval for the median of the estimator (which is the same as the mean if the distribution is symmetric) with a confidence coefficient of $1 - (\frac{1}{2})^{k-1}$ irrespective of the distribution of the estimator. It may be noted that the interpenetrating sub-samples are of value if the survey has to be carried out in successive stages due to the necessity of providing preliminary results. The agreement of the sub-sample estimates is likely to be more convincing to the layman than any statement of sampling and non-sampling errors.

Suppose there are two agencies and two parties of investigators within each agency to conduct the survey. Then 8 or a multiple of 8 (say 8k) independent interpenetrating sub-samples may be selected and each party of investigators in each agency may be assigned 2k

sub-samples at random for being surveyed. With this arrangement the total variation of the estimator may be analysed as given below.

source of variation	degrees of freedom
between agencies	1
between parties	2
within error	8k-4
total	8k-1

This analysis will help in locating the stages of operation where there is much of discrepancy. For instance if the between agency difference turned out to be statistically significant, this would mean that the survey has not been carried out according to the specifications by one or both the agencies. Similarly a significant result for the parties will help in locating that party which is not functioning according to the specifications.

12. ILLUSTRATIVE EXAMPLES.

An example of a situation where a sample survey estimate turned out to be nearer the true value than the complete enumeration figure, is provided by the Jute Survey in Bengal (India and Pakistan) during the years 1944-45 and 1945-46 (Mahalanobis and Lahiri, 1960). Jute being a cash crop of international importance, accurate figures for production become available subsequently. The official forecast in these years were based on complete enumeration of all plots. Sample

surveys were also conducted by the method of actual physical observations of randomly selected plots. The results of the enquiry are given in Table 2.

Table 2. Comparison of official (complete enumeration) and sample survey estimates with very reliable trade figures, Bengal 1944-45, 1945-46.

sr. no.	item	quantity (thousand bales) 1944-45	1945-46
(0)	(1)	(2)	(3)
1	trade figure	6728	7562
2	complete enumeration	4895	6304
3	sample survey	6480	7540
4	discrepancy between (2) and (1)	-27.2 percent	-16.6 percent
5	discrepancy between (3) and (1)	-3.6 percent	-0.3 percent

(Source : Mahalanobis, P.C. and Lahiri, D.B. (1960) 'Analysis of errors in censuses and surveys with special reference to experience in India', 32nd Session of the International Statistical Institute, Tokyo).

This interesting example shows that the sample survey provided a more accurate figure than the census because of the reduction in non-sampling errors made possible by confining the survey to a sample.

Another interesting case of response bias is provided by a study conducted in Central Iowa, USA by the Iowa State College (Hendricks, 1956). In this study the figures for volume of corn in

a sample of 50 cribs arrived at on the basis of farmers' judgement estimates were compared with the objective measurements got after the harvest. The result of this study showed that the judgement estimates was about 15 percent below the objective figures. The relevant figures regarding this study are given in Table 3.

The post enumeration survey in the 1950 Census of Population in the USA showed an under-enumeration of 1.4 percent which may be taken as the non-sampling bias in that census. Similarly the post-enumeration survey in the 1951 Census of Population in India showed an under-enumeration of 1.1 percent and the non-sampling bias in the 1956 Livestock Census in India was assessed to be about 15 percent (including processing errors) for large heads by a post-enumeration survey.

Table 3. Objective corn yield estimates compared with estimates from reported data.

area sampled	objective estimate		reported estimate	
	corn in field	adjusted for loss in harvest	unadjusted	adjusted to net acreage and 15 percent understatement
(1)	(2)	(3)	(4)	(5)
Alabama (1948)	26	23.4	21.0[x]	24.8
North Carolina (1949)	41	36.9	31.5[x]	37.2
Virginia (1949)	55	49.5	42.0[x]	49.6
10 Southern States	21.8	19.6	16.4[x]	19.4
Central Iowa (1953)	79.8	71.4	58.3[y]	68.8
" (1954)	74.0	66.6	55.7[y]	65.7

x - official estimates, y - reported data.
(Source : Hendricks, W.A. (1956) 'Non-sampling errors in agricultural surveys', Improving the quality of Statistical Surveys, 31-39, publication of American Statistical Association).

Mahalanobis and Lahiri (1960) have given an interesting example which brings out the utility of using interpenetrating sub-samples to serve as a broad check on the survey results. This refers to the Land Holding Survey conducted as a part of the National Sample Survey in India in 1954-55. In this survey the entire sample was drawn in the form of 12 independent and interpenetrating sub-samples, 8 of which were canvassed by the State agency and the other 4 sub-samples were surveyed by the Central agency. It may be noted that both the agencies carried out the survey using the same concepts and definitions, schedules and instructions, and the same programme of work. The results of this survey for some characteristics are presented in Table 4.

It may be noted from Table 4 that the Central and State estimates are not always in agreement with each other. It is found that the State estimates are lower than the Central estimates for the characteristics considered here. The difference between the Central and the State estimates are significant in case of number of households, number of operational holdings, land operated and land leased in.

Table 4. Showing the comparison between the Central and the State sub-sample estimates for aggregates of some characteristics
(estimates in millions)

sub-sample	category					
	no. of house-holds	no. of persons	no. of opera-tional holding	acreage operated	acreage owned	acreage leased in
1	61.0	298.7	59.5	294.7	270.7	24.0
2	58.8	285.3	58.0	323.3	274.0	49.0
3	59.9	292.8	59.4	313.6	268.4	45.2
4	56.5	280.1	55.7	290.4	257.4	33.0
			State sample			
1	56.3	286.7	54.8	278.2	254.9	23.3
2	56.7	283.5	54.9	281.0	248.4	32.6
3	58.3	295.0	57.2	269.9	246.1	23.8
4	56.8	280.6	48.4	301.6	276.3	25.3
5	58.8	301.1	57.3	300.4	269.9	30.5
6	57.2	285.3	55.7	289.1	271.0	18.1
7	56.1	285.0	53.6	300.0	271.9	28.1
8	56.1	280.2	48.6	277.5	270.5	7.0
			Pooled estimate			
central	59.0	289.2	58.1	305.4	267.6	37.8
State	57.1	287.2	53.8	287.2	263.6	23.6
			Difference			
actual	-1.98	-2.05	-4.33	-18.23	-4.02	-14.2
percen-tage	-3.4	- .07	-7.4	- 6.0	-1.5	-37.6
Student's t	2.43	.44	2.30[x]	2.23[x]	.61	2.51[x]

x - significant at 5 percent level

(Source : Mahalanobis, P.C. and Lahiri, D.B. (1960) 'Analysis of errors in censuses and surveys with special reference to experience in India', 32nd Session of the International Statistical Insitute, Tokyo).

13. USE OF QUALITY CONTROL TECHNIQUES.

The technique of statistical quality control (SQC) may be applied to census and survey work to assess the quality of the work and to improve the out-going quality with suitable corrective action. For this purpose it is desirable to those SQC techniques which have built in devices for initiating corrective action. More attention is to be paid to control of errors through SQC techniques than to acceptance plans for finished work. For a particular situation, the best plan is defined as that which ensures the highest out-going quality for a given cost or alternatively the lowest cost for a specified out-going quality. There is considerable scope to apply SQC techniques for control of errors in censuses and surveys because of the large amount of routine repetitive operations involved such as coding, punching etc.

No attempt will be made here to describe all the SQC techniques which may be applied to control errors in surveys. Instead one procedure is described which is indicative of such applications. Suppose k operators are ~~XX~~ doing a particular routine operation where the output can be checked and the permissible error rate in the finished work is specified. The work of each operator is first completely checked for a suitable length of time. If the error rate is less than the specified rate, only a sample of his work is verified in the subsequent periods of time. The decision ~~xx~~ regarding whether to continue verification on a sample basis or to have complete verification is taken separately for each operator on the basis of his cumulated error rate

over the past period. It may be noted that this procedure will help considerably in reducing the cost of verification and at the same time will ensure a specified quality level for the finished work. It may be mentioned that this type of procedure is being used in the United States Bureau of the Census and that this has been found to be helpful in controlling errors in census and survey work.

REFERENCES

1. BIRNBAUM, Z.W. and SIRKEN, M.G. (1950) — Bias due to the non-availability in sampling surveys, Jour. Amer. Stat. Assn., 45, 98-111.

2. DURBIN, J. (1956) — Non-response and call-backs in surveys, Bull. Inter. Stat. Inst., 34(2), 72-86.

DEMING, W. E. (1944) — On errors in surveys, Amer. Soc. Rev., 9, 359-369.

4. DEMING, W. E. (1960) — Sampling Design in Business Research, John Wiley and Sons.

5. EL BADRY, M.A. (1956) — A sampling procedure for mailed questionnaires, Jour. Amer. Stat. Assn., 51, 209-227.

6. GALES, KATHLEEN and KENDALL, M.G. (1957) — An enquiry concerning interviewer variability, Jour. Roy. Stat. Soc. 120A, 121-147.

7. GHOSH, B. (1949) — Interpenetrating (net-work) samples, Bull. Cal. Stat. Assn., 2, 108-119.

8. GRAY, P.G. (1956) — Examples of interviewer variability taken from two sample surveys, Appl. Stat., 5, 73-85.

9. HANSEN, M.H. and HURWITZ, W.N. (1946) — The problem of non-response error in sample surveys, Jour. Amer. Stat. Assn., 41, 517-529.

10. HANSEN, M.H., HURWITZ, W.N. (1946) — The problem of non-response error in sample surveys, Jour. Amer. Stat. Assn., 41, 517-529.

10. HANSEN, M.H., HURWITZ, W.N. and BURSHAD, M.A. (1960) — Measurement of errors in censuses and surveys, presented at the Meeting of the International Statistical Institute held in Tokyo.

11. HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953) — Sample survey methods and theory, John Wiley and Sons, Vol. II, 280-325.

12. HANSEN, M.H., HURWITZ, W.N., MARKS, E.S. and MAULDING, W.P. (1951) — Response errors in surveys, Jour. Amer. Stat. Assn. 46, 146-190.

13. HANSEN, R.H. and MARKS, E.S. (1958) Influence of the interviewer on the accuracy of survey results, Jour. Amer. Stat. Assn. 53, 635-655.

14. KISH, L. and SLATER, C.W. (1960) Two studies of interview variance of socio-psychological variables, presented at the Annual Meeting of the Amer. Stat. Assn.

15. LAHIRI, D.B. (1957) Recent developments in the use of techniques for assessment of errors in national surveys in India, Bull. Inter. Stat. Inst., 36(2), 71-93.

16. LAHIRID D.B. (1957) Observations on the use of interpenetrating samples in India, Bull. Inter. Stat. Inst., 36(3), 144-152.

17. MACURA, M. and BALABAN, V. (1960) Yugoslavian experience in evaluation of population censuses and sampling, presented at the Meeting of the Inter. Stat. Inst. held at Tokyo.

18. MAHALANOBIS, P.C. (1940) A sample survey of the acreage under Jute in Bengal, Sankhya, 4, 511-530.

19. MAHALANOBIS, P.C. (1944) On large scale sample surveys, Phil. Trans. Roy. Soc., 231 B, 329-451.

20. MAHALANOBIS, P.C. (1946) Recent experiments in statistical sampling in the Indian Statistical Institute, Jour. Roy. Stat. Soc., 109, 325-370.

21. MAHALANOBIS, P.C. (1950) Cost and accuracy of results in sampling and complete enumeration, Bull. Inter. Stat. Inst., 32(2), 210-213.

22. MAHALANOBIS, P.C. (1956) Statistics must have purpose, Presidential Address at Pakistan Statistical Conference.

23. MAHALANOBIS, P.C. and LAHIRI, D.B. (1960) Analysis of errors in censuses and surveys with special reference to experience in India, presented at the meeting of the Inter. Stat. Inst. held at Tokyo.

24. MURTHY, M.N. (1961) — Variance and confidence interval estimation, to be published in Sankhya.

25. MURTHY, M.N. and NANJAMMA, N.S. (1959) — Almost unbiased ratio estimates based on interpenetrating sub-sample estimates, Sankhya, 21, 381-392.

26. POLITZ, A.N. and SIMMONS, W.R. (1949) — An attempt to get the not-at-homes into the sample without call-backs, Jour. Amer. Stat. Assn., 44, 9-31.

27. SUKHATME, P.V. (1953) — Sampling Theory of Surveys with Applications, Iowa State College Press, 444-485.

28. SUKHATME, P.V. and SETH, G.R. (152) — Non-sampling errors in surveys, Jour. Ind. Soc. Agr. Stat., 4, 5-41.