Report of Progress of Cooperative Projects

of

The Institute of Statistics, North Carolina State College

The Agricultural Marketing Service

United States Department of Agriculture

February, 1956 - July, 1956

Progress Report No. 20 - 2
Mimeo Series #153

Comparison of Alternative Measures of Size in the Construction of Area Sampling Frames

by

Bernard S. Pasternack

TABLE OF CONTENTS

LIST OF TABLES

Chapter I

FRAME CONSTRUCTION AND AREA SAMPLING

1.1 Introduction

In attempting to construct area sampling frames with sampling units as close to equal size as possible, we are often faced with the problem of choosing from among various sources of information a specific "measure of size" to be used for the assignment of sampling units to the area segments which comprise the area universe. In general, each unit consists of a cluster of elements and the size of the sampling unit is defined to be the number of elements which it contains. Since the procedure for establishing an area sampling frame is based upon an "indicator" of size rather than on the actual size of these area segments, the variation in size among sampling units will depend upon the relationship of the "measure of size" we adopt to the actual size. It is obvious that variations in size of the sampling units (or clusters) may have a serious effect on the variance of the estimates of totals (or means per cluster) although moderate variation in size will usually have a relatively small effect on the variance of estimates of averages per element or ratios, Hansen and Hurwitz (1953).

The purpose of this thesis is to explore the relative merits of alternative measures of size in the construction of open country area sampling frames using data collected in North Carolina. However, before discussing the problem of the thesis in detail, a brief summarization of some of the fundamental aspects of frame construction and area sampling will be given. Perhaps this will provide a clearer insight into the significant role assumed by the chosen measure of size in producing variation in actual size between sampling units with consequently increased variability of the sampling estimates of totals. An excellent account of frame construction for area samples is given by Jessen and Thompson (1953).

In order to define the word frame in its restricted statistical sense, we will first offer an explanation of two commonly used words which are basic to sampling: universe and population. A finite set of sampling units defines the concept of a universe. The corresponding set of measures generated by the universe when an attribute or measurable characteristic is attached to each sampling unit defines the concept of a population. A frame is then defined to be a system which unambiguously determines all sampling units which comprise the universe under consideration.

Often, sampling statisticians are confronted with a number of different universes which are troublesome to sample due to the presence of one or both of the following circumstances: (i) the total number of elements in the universe is unknown (note that a sampling unit can consist of either one or a cluster of elements to be observed) or (ii) no specific designation scheme or frame exists by which to identify each element (or sampling unit). For example, it is known that, say in Union County, North Carolina, there exists a number of farms. Suppose it was desired to examine certain farm characteristics as they exist currently by means of a sample. The obvious first step would be to get a list of the farms in Union County and then to draw a random sample from the list following established procedures. However, a current list is not available. Hence there is no frame for farms in Union County, and therefore, a random sample of farms cannot be directly selected. This situation, in general, exists for the universes of persons, dwelling units, business establishments, etc.

1.2 The Area Method of Sampling

In order to circumvent this troublesome aspect, a procedure known as the "area method of sampling" has been developed so that probability samples of such universes can be drawn simply, yet with accuracy. As an illustration of the technique of area sampling consider the following example. It is desired to select for study a

probability sample from the universe of farms located in the open country portion of Union County, or in other words, the area of land exclusive of towns and cities. No list of these farms exists and only an approximate estimate of their total number (based on the most recent census figures) is available. A frame for this kind of universe does not exist because the farms are not identifiable in advance and their exact total number is unknown.

It is possible, nevertheless, to devise another universe of sampling units consisting of area segments, covering the open country portion of Union County, and arbitrarily constructed by the use of identifiable geographical boundaries such as roads, streams and railroads. Some method has to be employed to associate each farm with one and only one area segment or sampling unit. After this has been achieved, the number of farms or some other measure associated with any area segment can be regarded as a characteristic of that particular sampling unit. A probability sample of farms can thus be drawn by selecting a sample of area segments where the probability of choosing any farm will be equal to the probability of selecting the area segment with which it is associated. Even though the exact number and locations of farms is unknown, it is possible, by the use of this new universe of area segments, to obtain with known probabilities, a sample of farms. This example serves to illustrate how a universe unsuitable for statistical treatment can be transformed, by means of the area approach, into a new universe for which a frame actually exists.

1.3 Construction of an Area Sampling Frame

By regarding each area segment as the actual sampling unit, in the example cited above, a frame of, say $N = 395$, sampling units is obtained which includes all the farms in the universe of inquiry (open country portion of Union County). Each of these area segments can be numbered from 1 to 395 for identification purposes.

Identification of selected sampling units is expedited, however, if the open country portion of Union County is divided or partitioned into a number of large areas to be called divisions. Within each of these divisions every area segment can be numbered beginning with one until all have been enumerated. Thus if the open country portion of Union County is partitioned into, say 7 divisions, each area segment can be easily identified, first by its division number and second by its number within the division.

In the area universe presented above the area segment was arbitrarily chosen to be the sampling unit. Hence this universe consisting of 395 area segments forms a suitable frame for sampling since (i) the total number of sampling units is known and (ii) each can be uniquely determined by some simple procedure of numbering. Although a random sample could be adequately drawn from this area universe by means of the frame which has been constructed as above, such a universe may possess the undesirable property of having a large unit to unit variability for the measurable characteristic under examination. The problem of constructing the area sampling frame should also be governed by the extent to which a low sampling variance can be obtained.

If the number of farms actually varies considerably from area segment to area segment, the variance of the estimate of, say, the total number of farms in the open country portion of Union County will be quite high. On the other hand, it should be evident that the more equal each area segment is in actual size (i.e. number of farms), the lower will be the variance of this estimate. In fact, if area segments or sampling units can be constructed so that each segment contained an equal number of farms, the variance of the estimate of the total number of farms in the open country portion of Union County would reduce to zero. This is practically impossible without perfect advance knowledge, in which case there would be no need to estimate the total number of farms, but it serves to demonstrate the point in question.

1.4 Construction of a Frame to Equalize the "Size" of Area Sampling Units Using a Given "Measure of Size"

Construction of easily identifiable area segments or sampling units can only be achieved to a limited extent if a one-one correspondence between area segments and sampling units is required. If the originally specified area segments (designated as count units hereafter) can be assigned one or more sampling units by means of information available concerning their relative sizes, considerably greater "size equalization" of sampling units is possible. At this stage it becomes apparent that the problem of frame construction revolves about the choice of a technique or an appropriate measure of size for each area segment (count unit) which will enable the ultimate sampling units of the area universe to be as closely equal in size as possible. When alternative measures of size are available to aid in the accomplishment of this goal, a decision has to be made as to which one will be used. In North Carolina, area sampling frames were constructed in the Department of Experimental Statistics at North Carolina State College, until recently, by means of the information provided by the Master Sample materials. These were prepared by the Statistical Laboratory of Iowa State College in cooperation with the Bureau of Agricultural Economics and the Bureau of the Census during the period 1943-1944. The Master Sample materials[1] consist of small areas or count units covering the entire United States. Each of these count units has associated with it, an indicated number of farms (INOF) and an indicated number of dwellings (INOD). These indicated numbers were obtained (for the open country) by an actual count of the culture shown on maps prepared by the various State Highway Commissions. These counts of INOF and INOD thus provide a measure of

---

[1] For a discussion of the development of the Master Sample of Agriculture, see King and Jessen (1945).

size for each of the delineated count units.

On the basis of this information an area sampling frame for the open country portion of Union County can be constructed by assigning to each count unit an appropriate number of sampling units /based, say, on the Indicated Number of Farms as given by the Master Sample materials, i.e. (INOF) ms/. The total number of sampling units which would be assigned to the open country portion of Union County would be arrived at by dividing the current census figures for number of farms in the area under consideration, by the expected average size of sampling unit desired. The nearest integral number would represent the total number of sampling units to assign the entire area universe. The subsequent assignment (discussed below) of these sampling units to each count unit is then based upon the specific measure of size used in the construction of the area sampling frame. Hence, it is to be observed that the measure of size serves as the indicator of the distribution or density of the sampling units among the count units for the area universe frame to be constructed such that the sampling units are approximately equal in actual size.

The Master Sample materials consisting of area segments will always remain complete whatever changes may occur in the course of time, although the accuracy of the supplementary information provided by the map counts of the number of farms and dwelling units will naturally become progressively worse. It should be pointed out that in a simple random area sample the estimate of a total will always be unbiased. However, if the degree of relationship between the actual sizes and the size measures as provided by the Master Sample map count data decreases over time, the sampling variance of the estimated total for a given related to size characteristic, obtained from samples of the same size repeatedly over time, can be expected to increase. It appears reasonable to believe that after a certain period of time has elapsed extensive revision of the Master Sample materials will be necessary in certain parts of the

country where further use of the information provided by the Master Sample might result in unduly large variances of sample estimates, Yates (1949). In the state of North Carolina such a revision[1] has recently been made and these new materials (to be called Current) serve as the basis for choosing a measure of size for the construction of area sampling frames within the state. These new materials consist of a more convenient and more easily identifiable set of count units with supplementary current map count information on the number of farms and the number of dwelling units within each count unit.

1.5 A Technique for Selecting a Simple Random Sample of Area Sampling Units

Methods of constructing a universe of areas with "equalized size", under a given measure of size, exist by which much of the labor of actually constructing an area universe of segments can be eliminated without any loss of precision in the final results. Consider the following method used for Union County where the total number of sampling units in the open country is to be 2,164, a figure arrived at through the use of current census information. If (INOD)cc (i.e. Indicated Number of Dwellings based on Current Materials) is adopted as the measure of size for the construction of the area sampling frame, the assignment of an integral number of sampling units (to be denoted as su's) to any count unit will be based on the rule that the cumulative total of su's assigned through any count unit will be equal to the integer closest to cumulative (INOD)cc divided by the average "indicated size" of a su.

The average indicated size of a sampling unit for the area universe if defined to be total (INOD)cc /total no. of su's, which in the present situation equals 6063/2164 = 2.8018.

---

1 By the Survey Operations Unit, Institute of Statistics, North Carolina State College, Raleigh, North Carolina.

The frame consists of 2,164 sampling units. A random number between 1 and 2164 is drawn. An indication of the count unit in which the sampling unit lies is determined by multiplying the selected random number by the average indicated size of a sampling unit (in this case 2.8018). For example, if the random number chosen was 423 then (423)(2.8018) = 1185 to the nearest integer, which indicates that the su belongs to count unit 2-2 (see table 1). The number of sampling units assigned through that count unit and the count unit immediately preceding it is then computed by dividing column (3), in table 1, by the average indicated size of the sampling unit, 2.8018, and then recording the value obtained to the nearest whole number in column (4). For the example, through count unit 2-2 we have 1188/2.8018 = 424, and through the preceding count unit 2-1 we have 1180/2.8018 = 421. Therefore 3 su's are assigned to count unit 2-2. The number of sampling units assigned to the count unit is then recorded in column (5) by taking the difference between the number of sampling units assigned through the selected count unit and the number of sampling units assigned through the count unit immediately preceding it. Thus the computations essential for the assignment of su's to the count units is confined to a subset determined by the random selection. Note that the first three columns are filled out completely and constitute the materials used repeatedly for different samples. Computations for columns (4) and (5) depend upon the sample drawn.

Once the count units in which the selected sampling unit lies is obtained, it is divided into a number of segments exactly equal to the number of sampling units assigned to it, 3 for the example. Each sampling unit is then ordered according to some predetermined rule or technique such that, theoretically, in any repeated sampling process its number would remain the same.

Table 1

CONSTRUCTION OF AN AREA SAMPLING FRAME

(1) Count Unit	(2) INOD	(3) Cum. INOD	(4) SU's assigned through	(5) SU's assigned to
1- 1	7	7		
2	18	25		
3	10	35		
...	...	...		
79	14	1151		
2- 1	29	1180	421	
2	8	1188	424	3
3	10	1198		
...	...	...		
39	27	1646		
...	...	...	...	...
7- 1	15	5217		
2	18	5235		
3	22	5257		
...	...	...		
62	7	6063		

There are a large number of situations arising in area sampling where it is not practicable or desirable to designate the ultimate sampling unit as a specific area. Suppose, as in the example, that the selected sampling unit belongs to a count unit to which three sampling units have been assigned. The rule states that the count unit should be segmented into three parts of approximately an equal number of indicated dwellings. Suppose, however, that suitable roads, streams or other identifiable lines do not exist for use as boundaries within the count unit for the construction of sampling units of about equal indicated size. In this event, greater inequalities in the sizes of the final sampling units may have to be accepted, or else a different kind of sampling unit may have to be chosen such as one which consists of a systematic numbering of the households that actually exist in the count unit. In other words,

the three sampling units could be designated in the following manner:

Sample Unit No.	Elements (households) Contained
(1)	1,4,7,10,...
(2)	2,5,8,11,...
(3)	3,6,9,12,...

where the numbering of the households is fixed by some rule such as "the most northeast household is number 1 and others follow in a clockwise direction". When the households are not situated on a perimeter the rule required for numbering may be more complex but it should be such that each household's number is unambiguously determined. If the sampling unit selected was No. 3 in this particular count unit the field investigator would locate the count unit area, proceed to the northeast corner and list the households taking observations on the 3rd, 6th, 9th, etc., until the count unit area had been completely canvassed.

1.6 Statement of the Problem

At this point we will restate that the objective of this thesis is to compare alternative measures of size in the construction of area sampling frames in North Carolina with specific attention given to the measures of size referred to earlier, viz., (INOF)ms, (INOD)ms, (INOF)cc, and (INOD)cc. The choice of a criterion of comparison has been constrained to the property which has also been discussed previously, viz., the effect of these alternative measures of size on the variance of the sample estimates.

Chapter II

PRESENT STATUS OF RESEARCH ON THE EFFECT AND CONTROL OF VARIATION IN SIZE

2.1 Introduction

The discussion thus far has been limited to sampling procedures wherein the sampling units are not individual elements, but rather clusters of elements. Often, however, cluster sampling may be inefficient due to similarity among elements within clusters. This, in effect, is saying that a high positive intra-cluster correlation exists. Cost considerations also make cluster sampling prohibitive at times. In such situations it is logical to take measurements on only a sample of elements from each cluster rather than to enumerate it completely.

In order to present a generalized treatment of a sampling estimate and its variance consider a two-stage stratified random sampling scheme where the population is divided into, say, S strata. The clusters of elements (households or farms as in the earlier illustration) form the primary sampling units (i.e. psu's) in the area universe. In each stratum there are M_i psu's ($i=1,2,\ldots,S$) and in the j^{th} psu of the i^{th} stratum there are N_{ij} households ($j=1, 2, \ldots, M_i$). The sampling procedure in each stratum will be carried out in two stages. At the first stage a predetermined number of primary sampling units are sampled in each stratum and then at the second stage, also, a predetermined number of households are sampled in each selected primary sampling unit. In the i^{th} stratum a sample of m_i out of M_i psu's are selected at random at the first stage, with equal probability at each selection and without replacement. In each psu selected, all households are listed and n_{ij} out of the N_{ij} listed households are selected at random with equal probability and without replacement and then enumerated.

Let x_{ijk} be a variate representing the measure or value attached to any characteristic of the k^{th} household ($k=1, 2,\ldots,n_{ij}$) of the j^{th} primary unit ($j = 1,2,\ldots,m_i$) in the ith stratum ($i=1,2,\ldots,S$).

The total, X, for any given characteristic is given by

$$X = \sum_{i=1}^{S} \sum_{j=1}^{M_i} \sum_{k=1}^{N_{ij}} x_{ijk}$$

and the best linear unbiased estimate, X', of X is

$$X' = \sum_{i=1}^{S} M_i/m_i \sum_{j=1}^{m_i} N_{ij}/n_{ij} \sum_{k=1}^{n_{ij}} x_{ijk}$$

M_i/m_i and N_{ij}/n_{ij} are known as first-stage and second-stage expansion factors, respectively, and their reciprocals are referred to as sampling fractions.

In order to obtain the variance of an estimate in two-stage sampling it is important to note that the process of taking expectations must be distinguished over the two stages of sampling. That is to say, it will be necessary to have two sets of "expected values", one of which corresponds to the sampling of primary units and the other which will arise in the sampling of secondary units within a particular primary unit. The variance of the estimate, X', on the premise that n_{ij} is a number predetermined by some rule for each of the M_i primary units is [cf. Deming(1950)]:

$$V(X') = \sum_{i=1}^{S} M_i^2 \left\{ \frac{\sigma_i^2}{m_i} \frac{M_i - m_i}{M_i - 1} + \frac{1}{m_i M_i} \sum_{j=1}^{M_i} N_{ij}^2 \frac{\sigma_{ij}^2}{n_{ij}} \frac{N_{ij} - n_{ij}}{N_{ij} - 1} \right\}$$

where

$$\sigma_i^2 = \sum_{j=1}^{M_i} \frac{(x_{ij} - u_i)^2}{M_i}$$

in which

$$x_{ij} = \sum_{k=1}^{N_{ij}} x_{ijk}$$

$$\mu_i = \sum_{j=1}^{M_i} \frac{x_{ij}}{M_i}$$

and $$\sigma_{ij}^2 = \sum_{k=1}^{N_{ij}} \frac{(x_{ijk} - \mu_{ij})^2}{N_{ij}}$$

where $$\mu_{ij} = \sum_{k=1}^{N_{ij}} \frac{x_{ijk}}{N_{ij}} = \frac{x_{ij}}{N_{ij}}$$

For the case in which there are no strata so that m primary units are drawn from a universe consisting of M primary units, and then random subsamples are drawn consisting of n_j secondary units from the primary unit that was drawn at the j^{th} draw, the varianc of X' becomes

$$V(X') = \left(\frac{M}{m}\right)^2 \left\{\frac{M-m}{M-1} \quad m\sigma^2 + \frac{m}{M} \sum_{j=1}^{M} \left(\frac{N_j}{n_j}\right)^2 \quad \frac{N_j - n_j}{N_j - 1} \quad n_j \sigma_j^2\right\}$$

2.2 The Effect of Variation in Size

Hansen and Hurwitz (1943) in their classical paper on "The Theory of Sampling From Finite Populations" present an elegant analysis of the effect of change in size of first-stage units on the variance of the mean of a two-stage sample consisting of m first-stage units and n second-stage units per first-stage unit. Their analysis is based on the fact that they are able to express the variance of a mean as follows:

$$V(\bar{X}') = \frac{\sigma_0^2}{mn} \left[1 - \frac{n(m-1)}{N(M-1)} + \rho_1 \quad \left\{\frac{M-m}{M-1} \quad \frac{n}{N}(N-1) - \frac{N-n}{N}\right\}\right]$$

where ρ_1 represents the intra-class correlation coefficient within first-stage units of size N,

where $$\sigma_0^2 = \sum_{j=1}^{M} \sum_{k=1}^{N} \frac{(x_{jk} - \mu)^2}{MN}$$

and $$\rho_1 = \sum_{j=1}^{M} \sum_{k \neq k'} \frac{(x_{jk} - \mu)(x_{jk'} - \mu)}{\sigma_0^2 \quad MN(N-1)}$$

If the first-stage units are combined to give $\frac{M}{C}$ new first-stage units with CN second-stage units each, the variance of the mean (denoted now by $\bar{X}''$) of a two-stage sample of size mn will be given by

$$V(\bar{X}'') = \frac{\sigma_0^2}{mn}\left[1 - \frac{n(m-1)}{N(M-C)} + \rho_2\left\{\frac{M-mC}{M-C}\,\frac{n}{NC}\,(NC-1) - \frac{NC-n}{NC}\right\}\right]$$

where ρ_2 will now represent the intra-class correlation coefficient within first-stage units of size NC. The difference between the two variances can be expressed as

$$V(\bar{X}') - V(\bar{X}'') = \frac{\sigma_0^2}{mn}\left[\frac{n(m-1)}{N(M-1)}\,\frac{(C-1)}{(M-C)} + \alpha_1\rho_1 - \alpha_2\rho_2\right]$$

where

$$\alpha_1 = \frac{M-m}{M-1}\,\frac{n}{N}\,(N-1) - \frac{N-n}{N}$$

$$\alpha_2 = \frac{M-mC}{M-C}\,\frac{n}{NC}\,(NC-1) - \frac{NC-n}{NC}$$

Now since

$$\alpha_1 - \alpha_2 = \frac{n}{M}\left\{\frac{(C-1)(m-1)(MN-1)}{(M-1)(M-C)}\right\} \geq 0$$

and

$$\frac{n(m-1)(C-1)}{N(N-1)(M-C)} \geq 0$$

The conclusion is that

$$V(\bar{X}') - V(\bar{X}'') \geq 0$$

whenever $\rho_1 > \rho_2$ provided both ρ_1 and ρ_2 are positive. In other words, a gain in precision is brought about by enlarging first-stage units whenever the intra-class correlation is positive and decreases as the size of the first-stage unit increases. It also follows that the smaller the value of ρ_2 the larger is the gain, so that by choosing for consolidation those first-stage units which are as different as possible the gain can be increased. Note, though, that practical considerations put a limit on the size to which the first-stage units can be increased since cost of subsampling increases with larger and larger areas. Hence the increase in precision is to be weighed against the increase in cost. As an example, Sukhatme (1953) states that in

crop surveys, the variance is decreased when an administrative circle comprising a group of villages is used in place of a village as the first-stage unit of sampling, but practical considerations of cost and administrative convenience favor the use of the village.

If cost were no consideration the enlargement of first-stage units could proceed to a point of elimination of the use of first-stage units altogether and the second-stage units would be selected independently from the whole population.

In the more general situation of two-stage stratified sampling where it is assumed that the parameters of the frame M_i, N_{ij}, σ_i^2 and σ_{ij}^2 for each of the characteristics are fixed, the Hansen and Hurwitz development for examining the effect of variation in size of first-stage units cannot be extended along the same lines previously discussed. Very recently, however, J.C.Koop (1955) devised an ingenious approach to the development of the variance of an estimate in two-stage stratified sampling which is based upon an alternative two-stage sampling formulation. The variance formula he derives has certain advantageous properties with respect to frame construction which will subsequently be pointed out.

The specific values which M_i and N_{ij} assume in each stratum depend on the method of frame construction or the measure of size which is used. Usually the sampling system is such that M_i is always known, but N_{ij} is known only for those psu's which are selected at the first-stage of sampling. When M_i and n_{ij} are fixed in advance, it can be seen that the variance of the estimate of a total or mean will depend on the population values of these parameters with respect to a given frame.

In the technique for drawing a simple random sample of psu's illustrated at the end of the first chapter, the assignment of a number of psu's to a particular count unit or area segment depended solely upon the measure of size used in the construction of the area sampling frame. If an area universe is conceived of where each stratum is composed of a finite set of identifiable area segments, the choice of a measure of size

can be regarded as defining the allocation of primary units to each area segment. Each measure of size, in a sense, generates a configuration of primary units on to the area universe. If at a given point of time $\sum_{j=1}^{M_i} N_{ij}$ is constant for each stratum, then specific to each measure of size in the construction of the area sampling frame, M_i, N_i, σ_i^2 and σ_{ij}^2 will assume certain values. For the alternative measures of size to be compared in this thesis, the alternative frames which will be constructed will differ from each other in that each time a new measure of size is introduced a partial or total alteration of primary units will occur as some may be made larger and others smaller in size.

Consider a frame in which M_i is also held constant. If another frame is constructed still keeping M_i constant, but now rearranging the second-stage units, the σ_i^2's will change the σ_{ij}^2's will also change except for those primary units which are unaltered. The resulting configuration of primary units under the new frame does not change the mean value of N_{ij} but its variance changes depending upon the extent to which its frequency distribution is altered. Given that M_i is fixed in advance, a frame must exist for which the weighted sum of the σ_i^2's and σ_{ij}^2's will assume the lowest possible value for a given characteristic, assuming the practical limitations imposed by the relevant measures of size available for the construction of the frame.

For two-stage stratified sampling, Koop's derivation of the variance of the estimate of a total is based upon the postulate that a predetermined number of second-stage units are to be selected from each of the m_i primary units selected at the first stage independent of the actual sample of primary units. In other words, the psu that is selected first will n_1 su's drawn from it, the psu selected second will have n_2 su's drawn from it, etc. This differs from the Hansen and Hurwitz premise which, to repeat, is that the n_j secondary units chosen from the primary unit that was drawn on the j^{th} draw is a number predetermined by some rule for each of the M_i primary units.

The variance of the estimate of a total that Koop derives under his alternative two-stage sampling formulation is

$$V(X') = \sum_{i=1}^{S} N_i^2 \left\{ \frac{\sigma_i^2}{m_i} \frac{M_i - m_i}{M_i - 1} + \frac{1}{m_i n_{H_i} M_i} \sum_{j=1}^{M_i} N_{ij}^2 \sigma_{ij}^2 \frac{N_{ij} - n_{H_i}}{N_{ij} - 1} \right\}$$

where

$$n_{H_i} = \frac{m_i}{\sum_{p=1}^{m_i} 1/n_{i_p}}$$

and is therefore the harmonic mean of the sizes of the second-stage samples chosen from each of the m_i selected first-stage units in the stratum. As Koop observes, this result is of practical significance in the sense that, in any given stratum, once the prescribed sample sizes for the number of second-stage units, to be taken from each of the selected first-stage units, are determined (either by optimum allocation theory or by other considerations), it is only necessary to take from each primary unit a uniform second-stage sample of a size almost equal to the harmonic mean of the separate second-stage sample sizes, to achieve approximately the same degree of precision. The harmonic mean, n_{H_i}, will not assume integer values and hence the variance of a given estimate for any stratum will be a little greater or a little less than that for the case which would have resulted, if second-stage samples of of prescribed sizes had been taken, depending upon whether the nearest integer chosen is above or below the harmonic mean.

In order to investigate the effect of the variation in size of psu's in any given stratum, Koop recast his variance formula along the lines adopted by Hansen and Hurwitz for the case of psu's of equal size discussed earlier. He has shown that the variation in size of psu's contributes both directly through the term N_{ij}^2, and indirectly, through covariation with the characteristics under study, to the sampling error. The resulting formula, for any given estimate of a total, contains the intra-class correlation between the variates under study and certain measures of

variation and covariation involving also the N_{ij}'s. In essence it is, to a certain extent, a further generalization of the Hansen and Hurwitz result to the case where the psu's are of unequal size. Under the imposed alternative Hansen and Hurwitz two-stage sampling formulation, a similar expression could not be obtained since the analysis becomes possible when the predetermined second-stage sample size for each of the M_i primary sampling units is constant.

2.3 The Control of Variation in Size

As mentioned previously, the control of variation in the size of a sampling unit may be of much greater importance for the problem of obtaining an estimate of the aggregate of some characteristic for the population on the basis of a sample, than when estimating an average per element, percentage, or other ratio from a sample. Hansen and Hurwitz (1953) have investigated various methods of reducing the effect of variation in size of sampling units on the variance of the estimate of a total.

Probably, the most obvious method of reducing the contribution of the variation in size of su's to the variance of a simple unbiased estimate of a total with cluster sampling is to define su's that have a small amount of variation in size. That is, if there are adequate resources available, such as detailed maps and aerial photographs plus supplementary count information, a measure of size can be developed which would project a particular configuration of su's on to the area universe so as to considerably limit their variation in size. The Master Sample materials afford such measures of size, and they are being used to construct area sampling frames by methods similar to that illustrated in Chapter 1.

Selection with probability proportional to size (pps) will often aid in the control of variation in size of su's. When single-stage sampling is done by pps, an unbiased estimate of the total, X, is obtained, by multiplying the characteristic total for each su by the reciprocal of its probability of inclusion in the sample, and then adding the results for the selected units. In other words, for a sample of

m su's drawn with replacement

$$X' = \frac{A}{m} \sum_{i=1}^{m} \frac{X_i}{A_i}$$

where X_i is the total for the i^{th} su in the sample, A_i is the measure of size of the i^{th} su in the sample and A is the aggregate measure of size for the population.

If fairly detailed maps and other supplementary materials are available, subsampling of whole compact segments is feasible. In this procedure an indicated average size of a segment is decided upon and then the measure of size of each psu is defined in terms of the number of area segments or subsampling units into which the psu will be divided. The actual procedure can be carried out in the same manner as illustrated at the end of Chapter 1, or, if the indicated average size is, say, 3.5 and the estimate of the number of elements in a particular sampling unit is 40, then 40/3.5 = 11 segments or su's can be assigned to the psu. A whole number of segments will always have to be assigned to a psu rather than a number involving a fraction. The number of segments into which the psu's are divided become the measures of size which serve as the basis for the sample selection.

If an area sampling frame is constructed in which there is a large variation in size of psu's, another method of reducing the contribution of variation in size to the variance of unbiased estimates of totals is to stratify the psu's into a number of size groups before the sample is drawn. Hence psu's can be selected by use of varying sampling fractions in the different size groups. The subsampling process can then proceed, perhaps, by using a uniform over-all sampling fraction.

Other methods which can be put to use in controlling the effect of variation in size include the technique of applying ratio or other regression estimates, assuming that other aspects of optimum design for estimating ratios have been approximated. Also, it is important to bear in mind that the presence of a few sampling units of very extreme size in the population might contribute significantly to the sampling variance of an estimate. In such situations precautions should be adopted in order

to reduce or eliminate their possible effects (e.g. large institutional populations, such as prisons, hospitals and hotels can be treated as a special class in sampling, if such populations are to be included in the population that is being sampled. Such institutions can be identified in an area and separate sampling within them can be provided for.)

For two-stage stratified sampling, J. C. Koop (1955) has derived some further results which throw light on the nature of the effect of variation in size of primary sampling units when shifting from a given frame to one which is believed to be more efficient due to a reduction in variation of size. These results are again an aftermath of the variance formula he developed based on his alternative two-stage sampling formulation. He considers a frame in which the sizes of the psu's are unequal and a corresponding frame in which the psu's are equal in size. In both these frames the total number of psu's and su's are assumed constant. In this situation the variance of any given characteristic specific to the latter frame is less than that specific to the former, if a certain set of inequalities on the intra-class correlation coefficients of the characteristic in question are simultaneously satisfied for the frame in which the psu's are equal. Thus if a set of these parameters for the former frame are known or can be estimated from a sample, inequalities can be obtained indicating the limits between which the intra-class correlation coefficients of the latter frame should lie for the estimate in question to have a smaller variance. The minimum sample size required such that the estimates of the parameters in the former frame will be sufficiently precise to justify any conclusions suggested by the set of inequalities has not as yet been investigated. Further work has to be done in that direction, but in any event, it has been demonstrated that mere reduction in variation in size of psu's may not result in a more efficient sampling estimate if the corresponding changes in the essential intra-class correlations do not satisfy certain conditions.

CHAPTER III

METHODS OF COMPARISON

3.1 Introduction

In the summer of 1955 the Institute of Statistics at Raleigh, North Carolina, conducted a simple random area sampling survey in the open country portion of eleven Southern Piedmont counties in North Carolina in order to arrive at estimates of various agricultural characteristics, particularly total cotton acreage. Also, in 1955 a two-stage stratified area sample survey was conducted by the Survey Operations Unit of the Institute of Statistics in order to investigate various household characteristics including buying practices, particularly for those households which had members that had shopped at least once in the city of Winston-Salem in the year prior to interview. These two surveys will be referred to, respectively, as the Cotton study and the Winston-Salem study. The materials used in the construction of the area sampling frames for both of these studies were the heretofore mentioned Current materials. It was thought that the Master Sample materials developed in 1943-1944 (from 1937-1938 culture in North Carolina, for example) were decidedly out of date. John Monroe of the Institute of Statistics has summarized the characteristics of these materials as follows:[1]

> (1) abandonment of township boundaries as count unit boundaries. All count units are bounded only by county lines, roads, streams, railroads and city limits. (2) divisions within counties formed by the major road network, rather than the township as used in the Master Sample of Agriculture material. In most counties, this delineation renders a pie-shaped effect. (3) delineation of the unincorporated places defined in the 1950 Census. Enumeration district maps for those places were obtained from the Bureau of the Census, enabling the delineation of the area on the highway maps and the exclusion of those areas from the open country count. Unincorporated areas not defined by the Census (those places under 1,000 in population) are in the open country portions. (4) aerial photo count in congested areas around cities. Collection of photos for such areas is made as requests for samples are received. Use of this technique improves the accuracy of the highway maps count considerably. (5) city and town maps are made as the cities are drawn in samples.

The specific measure of size which was used for both of these studies was (INOD) cc.

---

[1] The Institute of Statistics, A Record of Research: III

In order to compare (INOD)cc against (INOF)cc and to compare (INOD)cc against both (INOD)ms and (INOF)ms a procedure was developed which made use of these two probability area samples. In essence, it involves the computing of weights which are used to calculate estimates of observations for a corresponding sample had it been drawn under an alternative measure of size. These estimates are based on the actual observations recorded for the characteristics measured in the chosen sample. The explicit underlying assumptions are (i) for a specific comparison the area universe to be considered remains the same for both measures of size and (ii) if the same size probability sample had been drawn from the alternative measure of size frame, exactly the same group of count units would have been represented although the actual configuration of the sampling units may have been different. The first of these assumptions can be approximately satisfied by defining out initial area universe in such a manner as to exclude, as nearly as possible, such portions for which neither Current nor Master Sample materials are available. Once an approximate coincidence of the two area universes is established so that effectively there exists but one such universe, the second assumption is intuitively reasonable. Hence the initial problem is to estimate the measures which would have been observed for each sampling unit had the area sampling frame been constructed with reference to the alternative measure of size. After this has been accomplished particular quantities can be determined, such as estimated sampling variances, which would upon comparison best reflect the relative efficiency of one measure of size to another.

3.2 Proposed Technique

The initial problem was resolved by considering this approach. Suppose a su belonging to some particular count unit is randomly drawn. Assume, further, that this count unit can be partitioned into the number of su's assigned to it such that the measures (for any specific characteristic) are the same for each of the su's within the count unit. If this type of partitioning of a count unit can be done for all

count units, an unbiased estimate of the measure (for any characteristic) of a su belonging to any count unit and drawn from a frame constructed under an alternative measure of size can be obtained. The notation to be used below should not be confused with a somewhat similar notation used previously. M_i and N_i as used here are defined specifically for the derivation given below. If the number of su's assigned to the i^{th} count unit is denoted by N_i for the measure of size actually used in the selection of the sample, and if the number of su's assigned to the same i^{th} count unit is denoted by M_i for the alternative measure of size, then $(N_i/M_i)x_i$ is an unbiased estimate where x_i represents the actual observed value of the randomly chosen su which belongs to the i^{th} count unit. If, as rarely occurs, two or more of the selected su's belong to the i^{th} count unit, the estimate for each of these su's for the sample based on the alternative measure of size is $(N_i/M_i)(\sum_{i=1}^{n_i} x_i/n_i)$ where n_i denotes the number of su's which fall into the i^{th} count unit.

However, such a partitioning would be impossible to accomplish in practice and consequently a bias is introduced which tends to make the variance of the sampling estimate of a mean or total smaller for the alternative measure of size when the procedure for estimating the observations is as described above. This can be seen as follows:

Consider a large area that is partitioned into k area segments which are called count units. Also, assume that two sets of information or "measures of size" (N_i and M_i) are available for assigning a fixed number of su's, say N, to the set of k count units.

Schematically,

Count Unit		$I(N_i)$		$II(M_i)$
1		N_1		M_1
2		N_2		M_2
...		...		...
k		N_k		M_k
	Total	N	=	M

Let x_{ij} be the measure for any characteristic of the j^{th} su lying in the i^{th} count unit under scheme I. Let y_{ij} be the measure of the same characteristic for the j^{th} su contained in the i^{th} count unit under shheme II. Note that

$$X_i = \sum_{j=1}^{N_i} x_{ij} = Y_i = \sum_{j=1}^{M_i} y_{ij}$$

For a single su drawn at random

$$E(x_{ij}-\mu)^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_i} \frac{(x_{ij}-\mu)^2}{N} = \sigma_x^2$$

where

$$\mu = \sum_{i=1}^{k} \sum_{j=1}^{N_i} \frac{x_{ij}}{N} = \sum_{i=1}^{k} \frac{N_i}{N} \bar{x}_i \quad \text{and} \quad \bar{x}_i = \frac{X_i}{N_i}$$

We can write

$$\sigma_x^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_i} \frac{(x_{ij} - \frac{X_i}{N_i} + \frac{X_i}{N_i} - \mu)^2}{N}$$

$$= \frac{1}{N}\left\{\sum_i \sum_j (x_{ij} - \frac{X_i}{N_i})^2 + 2\sum_i \sum_j (x_{ij} - \bar{x}_i)(\bar{x}_i - \mu) + \sum_i \sum_j (\bar{x}_i - \mu)^2\right\}$$

$$= \frac{1}{N}\left\{\sum_{i=1}^{k} \sum_{j=1}^{} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^{k} N_i(\bar{x}_i - \mu)^2\right\}$$

$$(1)\quad \sigma_x^2 = \sum_{i=1}^{k} \frac{N_i}{N} \sum_{j=1}^{N_i} \frac{(x_{ij}-\bar{x}_i)}{N_i} + \sum_{i=1}^{k} \frac{N_i}{N} (\bar{x}_i - \mu)^2$$

Similarly,

$$(2)\quad \sigma_y^2 = \sum_{i=1}^{k} \frac{M_i}{M} \sum_{j=1}^{M_i} \frac{(y_{ij}-\bar{y}_i)^2}{M_i} + \sum_{i=1}^{k} \frac{M_i}{M} (\bar{y}_i - \mu)^2$$

If a simple random sample is drawn, scheme I would be preferred to scheme II if $\sigma_x^2 < \sigma_y^2$. Note that (1) can be written as

$$\sigma_x^2 = \sum_{i=1}^{k} \frac{N_i}{N} \sigma_{w_{ix}}^2 + \sigma_{bx}^2$$

and (2) can be written as

$$\sigma_y^2 = \sum_{i=1}^{k} \frac{M_i}{M} \sigma_{w_{iy}}^2 + \sigma_{by}^2$$

If the y_{ij}'s are estimated by the weighting procedure described earlier, a bias will be introduced in the estimate of σ_y^2 since the within count unit component of variance $\sigma_{w_{iy}}^2$, is underestimated. This bias can be regarded as insignificant if the between count unit component of variance dominates the expression. These results can be extended in a similar fashion to a two-stage stratified area sample.

In the comparison of measures of size using the new materials as opposed to the Master Sample materials certain minor difficulties occur. Since, as previously stated, the count unit boundaries under the Current materials have been considerably revised, a correspondence of count units for the Master Sample materials and the Current materials was not readily available. In order to obtain such a correspondence it was necessary to obtain equivalent area segments on both maps, which at times required the combining of certain count units on either the Master Sample map or the Current map, or both. In the Cotton study where (IN9D) cc was the measure of size used in the construction of the area sampling frame, parts of count units were occasionally combined on the Master Sample map to assure a precise equivalence with count units on the Current map.

For the Winston-Salem study, the area segments for the alternative measure of size, (INOD)ms, were so chosen as to coincide with the area segments (not necessarily count units) selected using (INOD)cc as the measure of size. In other words, if a count unit was delineated into a smaller segment containing the selected su, the matched area on the Master Sample map was made to correspond to the delineated segment rather than to the count unit on the Current map. In such cases where an alteration of a count unit was made, the number of su's assigned to this, in a sense, redefined count unit, was computed by dividing the total (INOD)ms in the particular redefined count unit, by the indicated average size of a sampling unit under (INOD)ms.

CHAPTER IV

CALCULATIONS AND RESULTS

4.1 The Cotton Study

The area universe is the open country portion of Crop Reporting District No.8, an area with eleven counties in the Southern Piedmont area of North Carolina where cotton is by far the predominant crop. The populations under investigation for the comparison of alternative measures of size are:

(1) Number of Milk Cows

(2) Number of Beef Cattle

(3) Number of Hogs and Pigs

(4) Number of Cotton Fields

(5) Cotton Acreage

(6) Number of Corn Fields

(7) Corn Acreage

(8) Number of Wheat Fields

(9) Wheat Acreage

The sampling unit consists of an area segment having an expectation of four cotton fields. A random sample of 125 su's was selected on the basis of the estimated number of cotton fields in the eleven counties considered. No stratification was used and the observations were recorded on the basis of the "closed segment" approach. This method consists of taking observations on tracts of land which are confined within the boundaries of the selected area segment. In the "farm headquarters" approach observations are taken on those tracts of land which are owned by farmers who have their headquarters in the chosen area segment. These tracts of land are thus not confined within the boundaries of the selected sampling unit.

In order to reconcile the Master Sample materials with the Current materials, the 3 su's which fell in Mecklenburg County were excluded from the comparisons as were 2 others in Richmond County and 1 in Cabarrus County. There su's were eliminated in the process of redefining the area universes so as to be approximately the same

for both the old and new materials. All remaining 119 su's were used in the comparison of (INOD)cc vs. (INOD)ms and (INOD)cc vs. (INOF)ms, but out of these only 53 were used for the comparison of (INOD)cc vs. (INOF)cc. This was due to the fact that the (INOF)cc measure of size was available for just Union and Cleveland counties at the time of this study.

The variance of a mean is given by $(N-n/N-1)\ \sigma^2/n$ for a simple random sample. The statistic

$$s_x^2 = \sum_{i=1}^{n} (x_i-\bar{x})^2/n-1 \text{ and } s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{N_i}{M_i} x_i - \frac{1}{n} \sum_{i=1}^{n} \frac{N_i x_i}{M_i} \right]^2$$

was computed for the 9 characteristics under examination. The expected value of s_x^2 is, of course $(N/N-1)\sigma_x^2$. The relative efficiency of (INOD)cc was then calculated for each of the comparisons. The results of (INOD)cc vs (INOF)cc; (INOD)cc vs. (INOD)ms; and (INOD)cc vs. (INOF)ms are given in Tables 2, 3, and 4, respectively.

Table 2

COMPARISON OF ALTERNATIVE MEASURES OF SIZE: THE COTTON STUDY (n=53)

Characteristics	(INOD)cc s_x^2	(INOF)cc s_y^2	Relative Efficiency
Number of Milk Cows	350.6821	186.0112	53.043
Number of Beef Cattle	569.1517	924.9067	162.506
Number of Hogs and Pigs	145.1299	239.6976	165.161
Number of Cotton Fields	4.4528	5.6654	127.232
Cotton Acreage	105.4097	109.1335	103.533
Number of Corn Fields	3.8483	6.2487	162.376
Corn Acreage	55.2850	82.2569	148.787
Number of Wheat Fields	1.8229	3.1630	173.515
Wheat Acreage	123.2241	148.3140	120.361

Table 3

COMPARISON OF ALTERNATIVE MEASURES OF SIZE: THE COTTON STUDY (n = 119)

Characteristics	(INOD)cc s_x^2	(INOD)ms s_y^2	Relative Efficiency
Number of Milk Cows	320.9819	373.7680	116.445
Number of Beef Cattle	541.8302	610.8725	112.742
Number of Hogs and Pigs	930.1684	888.7839	95.551
Number of Cotton Fields	4.7754	3.9805	83.354
Cotton Acreage	111.8288	71.4165	63.862
Number of Corn Fields	24.0786	35.7756	148.578
Corn Acreage	258.1610	295.8185	114.587
Number of Wheat Fields	5.5266	4.7000	85.043
Wheat Acreage	154.7136	112.3320	72.606

Table 4

COMPARISON OF ALTERNATIVE MEASURES OF SIZE: THE COTTON STUDY (n=119)

Characteristics	(INOD)cc s_x^2	(INOF)ms s_y^2	Relative Efficiency
Number of Milk Cows	320.9819	245.4246	76.461
Number of Beef Cattle	541.8302	2373.6506	438.080
Number of Hogs and Pigs	730.1684	1832.3745	196.994
Number of Cotton Fields	4.7754	5.7675	120.775
Cotton Acreage	111.8288	81.5724	72.944
Number of Corn Fields	24.0786	36.1355	150.073
Corn Acreage	258.1610	274.1246	106.184
Number of wheat Fields	5.5266	3.8709	70.041
Wheat Acreage	154.7136	199.4962	128.945

The results of Table 1 and Table 3 tend to indicate that, in general, for an all purpose sample of agricultural characteristics using the closed segment approach (INOD) is a better measure of size for the construction of an area sampling frame than (INOF). This is particularly evident for (INOD)cc vs. (INOF)cc, although, there is only a slight gain in efficiency for the important characteristic of cotton acreage. For (INOD)cc vs. (INOF)ms there even appears to be a loss of efficiency in the estimation of cotton acreage.

Table 2 suggests that there is apparently not much gain in efficiency, if any, when (INOD)cc is the measure of size as opposed to (INOD)ms. In fact in this comparison there is, perhaps, a substantial loss in efficiency in the estimation of the cotton acreage. It should be borne in mind, however, that there is a bias operating in favor of the Master Sample materials although the assumption has been that it is of relatively small magnitude.

4.2 The Winston-Salem Study

The area universe covers the open country portion of 16 counties in the Northwest part of North Carolina. The populations under investigation for the comparison of (INOD)cc vs. (INOD)ms are:

(1) Number of persons under 12 years of age

(2) Number of persons 12-18 years of age

(3) Number of persons 18 years of age and older

(4) Number of households receiving a newspaper daily

(5) Number of households having a radio

(6) Number of households having a television set

(7) Number of households with income less than $3,000

(8) Number of households that shopped in Winston-Salem

(9) Number of white households

(10) Number of households

The psu consists of an area segment having an expectation of six households. The universe was stratified into 35 strata and $M_i = 500$ psu's were assigned to each of them. Every psu is composed of $N_{ij} = 2$ su's, and a two-stage stratified sample consisting of 105 su's was selected. This was accomplished by selecting $m_i = 3$ psu's from each stratum and $n_{ij} = 1$ su from each psu. The scheme of stratification and the number of psu's assigned to each county is presented in Table 5. Map count information for both (INOD)cc and (INCD)ms is also included.

In the process of redefining the area universe such that it would be approximately the same for the available information on both measures of size, Forsyth and Guilford counties were excluded from the comparison. Hence, 21 su's were eliminated from the study. Fourteen other su's distributed among the remaining counties were also removed from the study for the same reason. Thus the comparison of the two measures of size was based on a sample of 70 su's. This necessitated the introduction of a small bias in the estimate of the sampling variance since m_i was not equal to 3 for every stratum included in the sample.

The variance of the estimate of a total for a two stage stratified sample is (see page 12)

$$\sum_{i=1}^{S} M_i^2 \left\{ \frac{\sigma_i^2}{m_i} \frac{M_i - m_i}{M_i - 1} + \frac{1}{m_i M_i} \sum_{j=1}^{M_i} N_{ij}^2 \frac{\sigma_{ij}^2}{n_{ij}} \frac{N_{ij} - n_{ij}}{N_{ij} - 1} \right\}$$

which, for the sampling design for the Winston-Salem study, reduces to

$$\frac{M^2}{m} \sum_{i=1}^{S} \left\{ \frac{M-3}{M-1} \sigma_i^2 + \frac{1}{M} \sum_{j=1}^{M} 4\sigma_{ij}^2 \right\}$$

or

$$\frac{M^2}{m} \sum_{i=1}^{S} \left\{ \frac{M-3}{M-1} \sigma_i^2 + 4\sigma_{w_i}^2 \right\}$$

The statistic

$$s_b^2 = \sum_{i=1}^{S} s_{b_i}^2$$

Table 5

OPEN COUNTRY STRATIFICATION: THE WINSTON-SALEM STUDY

Country	No. of Strata	psu's assigned 500/Stratum	Map Count (INOD)cc	(INOD)cc per su	Map Count (INOD)ms	(INOD)Ms per su
Forsyth	4	2000	12374	6.19	2402	1.20
Davie	1	500	2742	5.48	1969	3.94
Rowan	3	1500	7483	4.99	4044	2.70
Davidson	3	1500	8024	5.35	4564	3.04
Randolph	3	1500	6891	4.59	3402	2.23
Guilford	3	1500	8267	5.51	4580	3.05
Rockingham	3	1500	6402	4.27	4260	2.84
Stokes	1.64	820	4342	5.30	2978	3.63
Surry	2.36	1180	6337	5.37	4224	3.58
Yadkin	1.48	740	4218	5.70	3297	4.46
Iredell	2.52	1260	6512	5.17	3962	3.14
Alexander	1	500	2827	5.65	1693	3.39
Wilkes	3	1500	7983	5.32	4803	3.20
Alleghany	.56	280	1807	6.45	1552	5.54
Ashe	1.44	720	3695	5.13	2599	3.61
Watauga	1	500	3191	6.38	2483	4.97

where

$$s^2_{b_{ix}} = \sum_{i=1}^{m_i} \frac{n_{ij}(\bar{x}_{ij}-\bar{x}_i)^2}{m_i-1}$$

was computed for each of the characteristics under examination for both measures of size (i.e., s^2_{bx} and s^2_{by} were computed). s^2_b represents the between primary unit within stratum mean square from an analysis of variance point of view. The expected value of $s^2_{b_i}$ is, in general, (under the Koop two-stage sampling formulation)

$$\frac{1}{M_i}\sum_{j=1}^{M_i}\sigma^2_{ij}\,\frac{N_{ij}}{N_{ij}-1} + \frac{\sum_{j=1}^{m_k} n^2_{ij} - n^2_i}{(m_i-1)n_i}\,\frac{1}{M_i}\sum_{j=1}^{M_i}\frac{\sigma^2_{ij}}{N_{ij}-1} + \frac{n^2_i - \sum_{j=1}^{m_i} n^2_{ij}}{(m_i-1)n_i}\,\frac{M_i}{M_i-1}\mu\sigma^2_i$$

where

$$n_i = \sum_{j=1}^{m_i} n_{ij} \quad \text{and} \quad \mu\sigma^2_i = \sum_{j=1}^{M_i}\frac{(\mu_{ij}-\bar{\mu}_i)^2}{M_i}$$

where

$$\bar{\mu}_i = \sum_{j=1}^{M_i}\frac{\mu_{ij}}{M_i}$$

This result is derived in Appendix B. In the Winston-Salem study (which satisfies both the Hansen and Hurwitz and Koop sampling formulations since $n_{ij} = 1$ and is constant) this reduces to

$$E\ s^2_b = \sum_{i=1}^{S}\frac{M}{M-1}\mu\sigma^2_i + \frac{1}{M}\sum_{j=1}^{M}\sigma^2_{ij}$$

or

$$E\ s^2_b = \sum_{i=1}^{S}\frac{M}{M-1}\mu\sigma^2_i + \bar{\sigma}_{w_i} = \frac{1}{4}\sum_{i=1}^{S}\frac{M}{M-1}\sigma^2_i + 4\bar{\sigma}_{w_i}$$

since when the N_{ij}'s are equal, say to N_i, $\mu\sigma^2_i = \frac{1}{N_i^2}\sigma^2_i$ as explained in Appendix B. In this case, $N_i = 2$ and hence the above formula.

Thus the statistic s^2_b was thought to be appropriate for the calculations of the relative efficiency of (INOD)cc vs. (INOD)ms since for the Winston-Salem study M M-3. The results of this comparison are given in Table 6.

Table 6

COMPARISON OF ALTERNATIVE MEASURES OF SIZE: THE WINSTON-SALEM STUDY (n=70)

Characteristics	(INOD)cc s^2_{bx}	(INOD)ms s^2_{by}	Relative Efficiency
Number of persons under 12 years of age	11.1706	15.2599	136.6077
Number of persons 12-18 years of age	3.4325	4.1119	119.7932
Number of persons 18 years of age and older	15.3254	38.9644	254.2472
Number of households receiving newspaper daily	1.8690	4.9046	262.4184
Number of households having a radio	4.0992	7.8179	190.7177
Number of households having a television set	1.9167	2.5675	133.9542
Number of households with income $3,000	1.5556	4.2919	275.9000
Number of households that shopped in Winston-Salem	2.2500	4.2529	189.0178
Number of white households	2.6984	6.8185	252.6868

The results of Table 6 clearly tend to indicate that for various household characteristics (INOD)cc is a superior measure of size for constructing an area sampling frame as opposed to (INOD)ms. Almost all of the 10 characteristics under study evidence a large gain in efficiency with the frame constructed on the basis of (INOD)cc.

Chapter V

SUMMARY AND CONCLUSIONS

The objective of this thesis is to compare alternative measures of size in the construction of area sampling frames in North Carolina. The specific measures of size under examination are (INOD)cc, (INOF)cc, (referring respectively to indicated number of dwellings and indicated number of farms as obtained from counts of current highway maps) and (INOD)ms, and (INOF)ms, (referring to counts provided by Master Sample materials). The criterion of comparison has been constrained to the relative effect of these alternative measures of size on the variance of sample estimates. In order to compare (INOD)cc against (INOF)cc and to compare (INOD)cc against both (INOD) and (INOF)ms a procedure was developed which made use of two probability area samples. The first of these was a study of agricultural characteristics in a sample of closed segments in the 8th Crop Reporting District of North Carolina and is referred to as the Cotton study. The second was a trade area survey of households in 16 counties surrounding Winston-Salem, North Carolina and is referred to as the Winston-Salem study. Both of these studies were drawn from frames constructed on the basis of (INOI The procedure, in essence, involves the computing of weights which are used to calculate estimates of observations for a corresponding (matched) sample had it been drawn under an alternative measure of size. These estimates are based on the actual observations recorded for the characteristics measured in the selected sample.

In the Cotton study the results of the comparison of alternative measures of size indicate that, in general, for an all purpose sample of agricultural characterist (INOD) is a better measure of size for the construction of an area sampling frame than (INOF). However, there is apparently little gain in efficiency, if any, when (INOD)cc is the measure of size as opposed to (INOD)ms. In fact in this comparison there is, perhaps, a substantial loss in efficiency for (INOD)cc in the estimation of total

cotton acreage. On the other hand, the results of the Winston-Salem study clearly suggest that for various household characteristics (INOD)cc is a superior measure of size for constructing an area sampling frame as opposed to (INOD)ms.

The inconclusive results of the Cotton study with respect to (INOD)cc vs. (INOD)m raises the question of whether or not the (INOD) measure of size is, in fact, an adequ source of information for developing a frame which is supposed to contain sampling units of approximately equal size with respect to number of cotton fields, particularl when the "closed segment" approach is used. It is quite possible that the assumption of a high degree of association between cotton fields and indicated number of dwelling is not justified. In any event, if some such relationship does exist it may be more beneficial to compare the alternative measures of size under the "farm headquarters" approach since the relationship between the measure of size and the sampling unit would be more meaningful in that situation.

This difficulty does not appear to the same extent in the Winston-Salem study since (INOD) can be surely expected to be an appropriate measure of size for construct- ing a frame "to equalize" the size of sampling units with respect to households. The results of this study lead to less tenuous conclusions. (INOD)cc demonstrated a large gain in efficiency for almost all of the 10 characteristics under examination.

The construction of sampling frames is of fundamental importance in sampling surveys and further investigations for determining the most suitable set of informa- tion available, with respect to a specific type of sampling inquiry, in the develop- ment of a frame should certainly be encouraged. Recent advances in the theory of frame construction should also be more fully explored, and when possible large scale sampling surveys should incorporate into the sampling design, frames that inherently provide a basis for the comparison of alternative measures of size.

APPENDIX A

Table 7

SAMPLING FRAMES FOR ALTERNATIVE MEASURES OF SIZE:

COTTON STUDY

Area Segment #	Sampling Units Assigned (INOD)cc	(INOF)cc	(INOD)ms	(INOF)ms
1	13	11	15	15
2	7	7	8	9
3	2	2	2	2
4	14	14	11	8
5	4	4	6	6
6	3	2	5	5
7	7	8	8	9
8	1	3	3	3
9	13	16	13	13
10	10	10	13	13
11	4	4	4	5
12	8	9	11	12
13	3	2	3	3
14	6	7	8	9
15	1	1	2	2
16	2	2	3	3
17	5	5	2	2
18	1	2	3	3
19	20	8	6	7
20	5	4	6	5
21	10	13	7	6
22	3	5	8	7
23	5	5	7	6
24	3	4	4	6
25	4	5	6	7
26	4	5	8	12
27	16	17	9	7
28	9	10	12	16
29	9	10	12	16
30	23	22	33	21
31	9	9	12	7
32	13	12	10	13
33	8	8	8	10
34	4	4	7	6
35	4	5	5	6
36	4	5	5	6
37	8	8	8	5
38	7	7	7	6
39	5	6	6	8
40	3	4	5	7

Area Segment #	Sampling Units Assigned (INOD)cc	(INOF)cc	(INOD)ms	(INOF)ms
41	3	1	3	2
42	21	21	19	22
43	9	9	10	10
44	14	14	12	10
45	5	3	5	4
46	6	6	7	6
47	11	10	10	10
48	11	9	11	11
49	10	8	11	10
50	2	3	1	1
51	6	5	8	6
52	4	2	4	4
53	12	11	11	9
54	12		10	10
55	6		6	6
56	4		4	6
57	14		11	12
58	9		7	7
59	18		22	23
60	5		6	6
61	11		12	12
62	8		9	10
63	8		12	11
64	9		11	10
65	7		3	3
66	2		1	1
67	36		19	22
68	19		16	15
69	3		11	5
70	6		3	1
71	8		9	5
72	5		8	6
73	5		4	5
74	2		3	3
75	6		6	6
76	6		7	7
77	8		3	1
78	2		3	4
79	4		4	5
80	3		3	4
81	4		3	3
82	5		7	8
83	2		2	3
84	2		3	4
85	2		4	4
86	2		3	4
87	3		3	3
88	3		3	4

Area Segment #	Sampling Units Assigned (INOD)cc	(INOF)cc	(INOD)ms	(INOF)ms
89	10		11	7
90	4		5	3
91	4		6	8
92	14		22	25
93	4		5	6
94	3		7	7
95	2		3	3
96	2		3	4
97	5		6	9
98	9		11	12
99	6		7	7
100	9		10	12
101	22		10	4
102	20		5	2
103	20		5	2
104	4		5	5
105	15		22	15
106	3		4	5
107	5		4	5
108	3		3	2
109	1		3	2
110	5		3	4
111	5		5	6
112	5		5	3
113	2		4	3
114	2		2	3
115	2		2	2
116	5		4	4
117	2		1	1
118	2		4	4
119	4		5	4

Table 8

SAMPLING FRAMES FOR ALTERNATIVE MEASURES OF SIZE:

WINSTON-SALEM STUDY

Area Segment #	SU's Assigned (INOD)cc	(INOD)ms	Area Segment #	SU's Assigned (INOD)cc	(INOD)ms
1	3	2	36	3	4
2	5	7	37	1	1
3	5	5	38	1	1
4	1	3	39	1	1
5	2	2	40	1	1
6	3	6	41	2	2
7	1	1	42	1	1
8	4	7	43	1	1
9	6	9	44	4	2
10	1	2	45	2	3
11	3	6	46	2	2
12	4	7	47	2	3
13	4	5	48	5	7
14	4	4	49	3	4
15	2	2	50	4	2
16	3	9	51	1	2
17	2	3	52	5	6
18	2	1	53	2	4
19	4	6	54	4	6
20	2	1	55	2	2
21	4	5	56	2	3
22	3	2	57	2	3
23	2	2	58	5	4
24	3	2	59	4	4
25	2	3	60	3	4
26	5	7	61	2	2
27	3	4	62	3	4
28	4	3	63	8	7
29	3	3	64	2	2
30	5	4	65	3	3
31	3	2	66	3	2
32	3	3	67	1	1
33	1	1	68	2	1
34	1	2	69	4	4
35	3	6	70	1	2

APPENDIX B

In the analysis of variance the following algebraic identity for the i^{th} stratum is well known. The notation has been kept consistent with that used in section 2.1.

$$\sum_{j,k} (x_{ijk}-\bar{x})^2 = \underbrace{\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} (x_{ijk}-\bar{x}_{ij})^2}_{W_i} + \underbrace{\sum_{j=1}^{m_i} n_{ij}\,(\bar{x}_{ij}-\bar{x}_i)^2}_{B_i}$$

where

$$\bar{x}_i = \frac{\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} x_{ijk}}{\sum_{j=1}^{m_i} n_{ij}} \quad \text{and} \quad \bar{x}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} x_{ijk}$$

Before proceeding any further let us define

$$\mu\sigma_i^2 = \sum_{j=1}^{M_i} \frac{(\mu_{ij}-\bar{\mu}_i)^2}{M_i} \quad \text{where} \quad \bar{\mu}_i = \sum_{j=1}^{M_i} \frac{\mu_{ij}}{M_i}$$

When all the primary units have the same number of second-stage units, say $N_{ij} = N_i$, it will be seen that $N_i^2\,\mu\sigma_i^2 = \sigma_i^2$ from the following considerations:

$$x_{ij} = \sum_{k=1}^{N_{ij}} x_{ijk} = \mu_{ij}N_i \quad \text{since } \mu_{ij} = \frac{\sum_{k=1}^{N_{ij}} x_{ijk}}{N_{ij}}$$

$$\sigma_i^2 = \sum_{j=1}^{M_i} \frac{(x_{ij}-\mu_i)^2}{M_i} = \sum_{j=1}^{M_i} \frac{(\mu_{ij}N_i - N_i\bar{\mu}_i)^2}{M}$$

since

$$\mu_i = \sum_{j=1}^{M_i} \frac{x_{ij}}{M_i} = \sum_{j=1}^{M_i} \frac{\mu_{ij}N_i}{M_i} = N_i \sum_{j=1}^{M_i} \frac{\mu_{ij}}{M_i} = N_i\bar{\mu}_i$$

Therefore,

$$\sigma_i^2 = N_i^2 \sum_{j=1}^{M_i} \frac{(\mu_{ij}-\bar{\mu}_i)^2}{M_i} = N_i^2 \mu c_i^2$$

We have

$$B_i = \sum_{j=1}^{m_i} n_{ij} (\bar{x}_{ij}-\bar{x}_i)^2$$

For convenience the subscript i will be dropped in the derivation of $E(B_i)$. Thus we have

$$B = \sum_{j=1}^{m} n_j (\bar{x}_j-\bar{x})^2$$

$$= \sum_{j=1}^{m} n_j \bar{x}_j^2 - n\bar{x}^2, \text{ where } n = \sum_{j=1}^{m} n_j$$

$$= \sum_{j=1}^{m} \frac{(\sum_{k=1}^{n_j} x_{jk})^2}{n_j} - \frac{1}{n} (\sum_{j=1}^{m} \sum_{k=1}^{n_j} x_{jk})^2$$

$$= \sum_{j=1}^{m} \frac{1}{n_j} (\sum_{k=1}^{n_j} x_{jk})^2 - \frac{1}{n}\left\{\sum_{j=1}^{m} (\sum_{k=1}^{n_j} x_{jk})^2 + \sum_{p\neq q} (\sum_{k=1}^{n_p} x_{pk})(\sum_{k=1}^{n_q} x_{qk})\right\}$$

$$E(B \mid j=1,2,\ldots,m) = \sum_{j=1}^{m} \frac{1}{n_j} E(\sum_{k=1}^{n_j} x_{jk})^2 - \frac{1}{n} E\left\{\sum_{j=1}^{m} (\sum_{k=1}^{n_j} x_{jk})^2 + \sum_{p\neq q} (\sum_{k=1}^{n_p} x_{pk})(\sum_{k=1}^{n_q} x_{qk})\right\}$$

Now $(\sum_{k=1}^{n_j} x_{jk})^2 = n_j^2 \bar{x}_j^2$

therefore,

$$E(\sum_{k=1}^{n_j} x_{jk})^2 = n_j^2 E(\bar{x}_j^2) = n_j^2 (V(\bar{x}_j) + \mu_j^2)$$

$$= n_j^2 \left\{\frac{\sigma_j^2}{n_j} \frac{N_j-n_j}{N_j-1} + \mu_j^2\right\} = n_j \sigma_j^2 \frac{N_j-n_j}{N_j-1} + n^2 \mu_j^2$$

(where it must be remembered that $N_j = N_{ij}$ and $\mu_{ij} = \mu_j$ in view of the fact that the subscript i has been dropped) so,

$$E(B\,|\,j=1,2,\ldots,m) = \sum_{j=1}^{m} \frac{1}{n_j}\left\{n_j\sigma_j^2\ \frac{N_j-n_j}{N_j-1} + n_j^2\mu_j^2\right\} - \frac{1}{n}\left[\sum_{j=1}^{m}\left\{n_j\sigma_j^2\ \frac{N_j-n_j}{N_j-1} + n_j^2\sigma_j^2\right\} + \sum_{p\neq q} n_p n_q \mu_p \mu_q\ \right]$$

$$= \sum_{j=1}^{m}\left\{\sigma_j^2\ \frac{N_j}{N_j-1} - \frac{\sigma_j^2}{N_j-1}\,n_j + n_j\mu_j^2\right\} - \frac{1}{n}\left[\sum_{j=1}^{m}\left\{\sigma_j^2\ \frac{N_j n_j}{N_j-1} - \frac{\sigma_j^2 n_j^2}{N_j-1} + n_j^2\mu_j^2\right\} + \sum_{p\neq q} n_p n_q \mu_p \mu_q\right]$$

$$E\left\{E(B\,|\,j=1,2,\ldots,m)\right\} = E\sum_{j=1}^{m}\left\{\sigma_j^2\ \frac{N_j}{N_j-1} - \frac{\sigma_j^2}{N_j-1}\,n_j + n_j\mu_j^2\right\}$$

$$- \frac{1}{n}E\left[\sum_{j=1}^{m}\left\{\sigma_j^2\ \frac{N_j n_j}{N_j-1} - \frac{\sigma_j^2 n_j^2}{N_j-1} + n_j^2\mu_j^2\right\} + \sum_{p\neq q} n_p n_q \mu_p \mu_q\ \right]$$

and on the basis of Koop's sampling formulation where the n_j are fixed in advance for each of the psu's sampled, we find

$$E\left\{E(B/j=1,2,\ldots,m)\right\} = \frac{m}{M}\sum_{j=1}^{M}\sigma_j^2\ \frac{N_j}{N_j-1} - \frac{1}{M}\sum_{j=1}^{M}\frac{\sigma_j^2}{N_j-1}\left\{\sum_{j=1}^{m} n_j\right\} + \frac{1}{M}\sum_{j=1}^{M}\mu_j^2\left\{\sum_{j=1}^{m} n_j\right\}$$

$$- \frac{1}{n}\left[\left\{\sum_{j=1}^{m} n_j\right\}\frac{1}{M}\sum_{j=1}^{M}\frac{\sigma_j^2 N_j}{N_j-1} - \left\{\sum_{j=1}^{m} n_j^2\right\}\frac{1}{M}\sum_{j=1}^{M}\frac{\sigma_j^2}{N_j-1} + \left\{\frac{1}{M}\sum_{j=1}^{M}\mu_j^2\right\}\left\{\sum_{j=1}^{m} n_j^2\right\}\right.$$

$$\left. + \left\{\sum_{p\neq q} n_p n_q\right\}\left\{-\frac{\mu\sigma^2}{M-1} + \bar{\mu}^2\right\}\right]$$

$$\text{(where } {}_{\mu}\sigma^2 = {}_{\mu}\sigma_i^2 \text{ and } \bar{\mu} = \bar{\mu}_i\text{)}$$

$$\text{i.e. } E(B) = (m-1)\ \frac{1}{M}\sum_{k=1}^{M}\frac{N_j}{N_j-1}\sigma_j^2 + \left\{\sum_{j=1}^{m}\frac{n_j^2-n^2}{n}\right\}\left\{\frac{1}{M}\sum_{j=1}^{M}\frac{\sigma_j^2}{N_j-1}\right\}$$

$$+\left(\mu\sigma^2+\mu^{-2}\right)\left\{\frac{n^2-\sum_{j=1}^{m}n_j^2}{n}\right\} + \frac{\mu\sigma^2}{n(M-1)}\sum_{p\neq q} n_p n_q - \frac{\mu^{-2}}{n}\sum_{p\neq q} n_p n_q$$

$$= (m-1)\ \frac{1}{M}\sum_{j=1}\frac{N_j}{N_j-1}\sigma_j^2 - \left\{\frac{n^2-\sum_{j=1}^{m}n_j^2}{n}\right\}\left\{\frac{1}{M}\sum_{j=1}^{M}\frac{\sigma_j^2}{N_j-1}\right\}$$

$$+\mu\sigma^2\left\{\frac{n^2-\sum_{j=1}^{m}n_j^2}{n}\right\}\frac{M}{M-1}$$

and

$$E(s_{b_i}^2) = E\left(\frac{B_i}{m_i-1}\right) = \frac{1}{M_i}\sum_{j=1}^{M}\frac{N_{ij}}{N_{ij}-1}\sigma_{ij}^2 - \frac{n_i^2-\sum_{j=1}^{m}n_{ij}^2}{n_i(m_i-1)}\left\{\frac{1}{M_i}\sum_{j=1}^{M_i}\frac{\sigma_{ij}^2}{N_{ij}-1}\right\}$$

$$+\mu\sigma_i^2\left\{\frac{n_i^2-\sum_{j=1}^{m}n_{ij}^2}{n_i(m_i-1)}\right\}\frac{M_i}{M_i-1}$$

with the subscript i included.

To complete the problem, it can similarly be shown that

$$E(W_i \mid j=1,2,\ldots,m) = \sum_{j=1}^{m_i}(n_{ij}-1)\,\sigma_{ij}^2\,\frac{N_{ij}}{N_{ij}-1}$$

so that

$$E(W_i) = E\left\{E(W_i \mid j=1,2,\ldots,m)\right\} = \left\{\frac{1}{M_i}\sum_{j=1}^{M_i}\sigma_{ij}^2\ \frac{N_{ij}}{N_{ij}-1}\right\}\left\{\sum_{j=1}^{m_i}(n_{ij}-1)\right\}$$

$$= \frac{1}{M_i}\sum_{j=1}^{M_i}\sigma_{ij}^2\ \frac{N_{ij}}{N_{ij}-1}\left\{(n_i-m_i)\right\}$$

$$\text{and } E(S_{W_i}^2) = E\left(\frac{W_i}{n_i-m_i}\right) = \frac{1}{M_i}\sum_{j=1}^{M_i}\sigma_{ij}^2\ \frac{N_{ij}}{N_{ij}-1}$$

BIBLIOGRAPHY

Deming, W. E. 1950. Some Theory of Sampling. John Wiley and Sons, Inc., New York.

Hansen, M. H., and Hurwitz, W. N. 1943. On the theory of sampling from finite populations. Ann. Math. Stat. 14:333-362.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. 1953. Sample Survey Methods and Theory: Vol. 1. John Wiley and Sons, Inc., New York.

Jessen, R. J., and Thompson, D. J. 1953. Design and Analysis of Surveys, Chapter 8. Dittoed Notes. Iowa State College, Ames, Iowa.

King, A. J., and Jessen, R. J. 1945. The master sample of agriculture. Jour. Amer. Stat. Assn. 40:38.

Koop, J. C. 1955. Sample Survey of Labor Force in Rangoon: A Study in Methods. Union Government Printing and Stationary, Rangoon, Burma.

Sukhatme, P. V. 1953. Sampling Theory of Surveys with Applications. Iowa State College Press, Ames, Iowa.

Yates, F. 1949. Sampling Methods for Censuses and Surveys. Charles Griffin and Co., London.