ON SAMPLING FOR STATE ESTIMATES OF THE FARM POPULATION IN NORTH CAROLINA USING THE TOWNSHIP AS A PRIMARY SAMPLING UNIT

by

Marvin A. Kastenbaum

A thesis submitted to the Faculty of the State College of Agriculture and Engineering of the University of North Carolina in partial fulfillment of the requirements for the degree of Master of Science in Experimental Statistics.

Raleigh, N. C.
1950
ACKNOWLEDGMENT

I wish to express my indebtedness to Dr. R. L. Anderson and Mr. A. L. Finkner for the constant guidance and generous assistance which they rendered during the course of this study.

Acknowledgment is also made to Mr. W. D. Lee of the Soils Department of the North Carolina State College for his technical advice regarding the delineation of the true soil strata in North Carolina.

I express my sincere thanks to Mr. J. S. Hunter and Mrs. K. Osbeck without whose efforts the data for this report might not have been tabulated.

M. A. K.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>II. Method of Stratification</td>
<td>2</td>
</tr>
<tr>
<td>III. Sampling Plan</td>
<td>6</td>
</tr>
<tr>
<td>IV. Calculations and Results</td>
<td>11</td>
</tr>
<tr>
<td>V. Appendix: Formulae and Notation</td>
<td>19</td>
</tr>
<tr>
<td>A. Unbiased Estimate</td>
<td>20</td>
</tr>
<tr>
<td>B. Ratio Estimate</td>
<td>24</td>
</tr>
<tr>
<td>C. Covariance of the Two Unbiased Estimates</td>
<td>26</td>
</tr>
<tr>
<td>VI. Bibliography</td>
<td>37</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table I. Comparison of Stratifications Based on Primary Sampling Units of Different Size ................. 13
Table II. Comparison of Four Different Types of Stratification for "Best" Overall Estimates ............... 17

LIST OF MAPS

Map I. Types of Farming Areas in North Carolina
        With County Boundaries Preserved .................. 3
Map II. Types of Farming Areas in North Carolina
        With Township Boundaries Preserved ............... 5
INTRODUCTION

In sampling for state estimates of characteristics of the farm population, it is desirable to obtain one estimate which is "best" for all agricultural and socio-economic characteristics under investigation. It has been found that a particular estimate which is "best" for one characteristic may not be very good for another. We therefore endeavor to find a sampling scheme which will yield a "best" overall estimate.

It is the purpose of this study to investigate the adequacy of four different methods of stratification for making estimates of certain agricultural and socio-economic characteristics of the farm population of North Carolina. We are not interested here in obtaining a "best" estimate for any particular item. It is rather our desire to obtain a sampling design which will yield the smallest variances for as many characteristics as possible.

The problem is considered in a paper by Madow (1). In this study, the author employs a two stage sampling design in which the county is the primary sampling unit, and the master sample segments, the sub-sampling units. A similar sampling plan is employed in our study. However, instead of the county, the minor civil division (township; M C D) is used as the primary sampling unit.

---

1/ The word "best" will be used throughout this paper to define an estimate whose coefficient of variation is smallest. It should be noted that no considerations are made on the basis of the cost of obtaining such an estimate. In practice, however, the smallest variance may not indicate the best estimate to be used, if the cost of obtaining such an estimate is too high. Furthermore, the bias of an estimate may effect the definition of "best". Inasmuch as the bias factor is negligible in our calculations, we have not considered it at all.

2/ The master sample segments are small areas of land with identifiable boundaries. The average number of farms per master sample segment in the United States is five. A full description of these units is given in a paper by Jessen and King (2).
The main questions under investigation are given as follows:

1. Will a geographic stratification with minor civil divisions as primary sampling units yield "better" estimates than a geographic stratification with the county as primary sampling unit? The figures which are used for the county as primary sampling units are those which appear in the paper by Madow (1). The problem of comparing the estimates derived in this report to those of Madow involves some difficulty because of the difference in the size of the primary sampling units. A further discussion of this problem is outlined below.

2. Will a stratification based on some socio-economic characteristic yield better estimates than a purely geographic stratification? Two methods of stratification are used as bases for comparison. These are stratifications by % non-white operators and by % telephones.

3. Will a stratification based on some agricultural characteristic yield better estimates than a purely geographic stratification? In this case a stratification based on % tobacco acreage was employed.

4. Which of the four methods of stratification under investigation yields the "best" overall estimates of the characteristics?

2. METHODS OF STRATIFICATION

The 100 counties in North Carolina are classified into twelve contiguous soil strata. These soil strata are used merely as guides for the true strata in this study. Thus, for example, in preparing the geographic stratification, the twelve major soil strata were outlined on map 1, and each was divided into smaller strata.

---

1/ See footnote 1 on page 1.

2/ Due to insufficient information at the time of this stratification, it was necessary to restrict the soil strata lines to the county boundaries; that is, no county boundary was crossed by a stratum line. Thus certain minor civil divisions which may have ordinarily been in one stratum were of necessity placed into another. This situation was however remedied for the other stratifications.
Map 1. Types of Farming Areas in North Carolina With County Boundaries Preserved

I. Northern Tidewater
II. Southern Tidewater
III. North Central Coastal Plain
IV. Central Coastal Plain
V. Upper Coastal Plain
VA. Sandhills

VB. Lower Piedmont
VI. Northern Piedmont
VII. Central Piedmont
VIIA. Foothills
VIII. North Mountain
VIIIA. Southern Mountain
containing several contiguous townships. Each new stratum was determined by totalling the 1940 number of farms within its contiguous townships until the number reached approximately 1400. The major soil strata lines were necessarily crossed in several instances to maintain geographic contiguity. Such a stratification based on the 1940 Census of Agriculture figures yielded exactly 197 strata with an average of 1404 farms per stratum. The smallest stratum contained 1226 farms; the largest contained 1645 farms.

It is to be noted that this particular geographic stratification is not unique. In fact any number of such stratifications are possible. The one used in this study was evolved by a trial and error method only.

After having completed the geographic stratification, it was agreed that a better definition of the twelve major soil strata was necessary. In an effort to define more exact stratum lines for the remaining stratifications, the author met on several occasions with W. D. Lee of the Soils Department at the North Carolina State College. With the aid of individual county soil maps, new stratum lines were mapped which were restricted only by the boundaries of the minor civil divisions (see map II).

Similar methods of stratification were used for the three remaining breakdowns. The minor civil divisions were listed within each major soil stratum in

---

1/ Two hundred was chosen as the desired number of strata in this sampling scheme so as to facilitate comparison with the twenty strata used by Madow (1). In sampling one county as a primary sampling unit from twenty strata, figures are ultimately obtained for approximately 200 minor civil divisions. In view of the fact that the sampling scheme in this study involves the MCD as the primary sampling unit, approximately 200 strata were necessary. Hence since there were 276,604 farms in North Carolina in 1940 - exclusive of Swain and Dare counties - the number of farms per stratum came to about 1400. Although equal size strata were constructed for this investigation, some authors see no need to impose such a restriction on the sampling scheme. Jebe (5) points out that some of his results, (C.V.)², obtained for North Carolina in terms of the between-county components of error for comparable items are less than or equal to those obtained by Madow (1) who used strata equalized as nearly as possible in size as measured by number of farms.
Map 2. Types of Farming Areas in North Carolina With Township Boundaries Preserved

I. Northern Tidewater
II. Southern Tidewater
III. North Central Coastal Plain
IV. Central Coastal Plain
V. Upper Coastal Plain
VA. Sandhills
VIB. Lower Piedmont
VI. Northern Piedmont
VII. Central Piedmont
VIIA. Foothills
VIII. North Mountain
VIII A. Southern Mountain
order of magnitude. Thus, for instance, a township with 5% tobacco acreage preceded one with 14% tobacco acreage. The formation of strata from these listing sheets depended only on the % tobacco acreage in the MCD. Again as in the geographic stratification, the 1940 number of farms within each MCD of about equal % tobacco acreage was totalled until that number reached approximately 1400. However, in this case, the major soil stratum lines were never crossed. Here each soil stratum was treated as a single unit, and the new strata were developed within each such unit. The same procedure was used for breakdowns by % non-white operators and % telephones. In each instance, 198 strata resulted.

The percentages by which the MCD's were stratified are evaluated in the following manner:

\[
\% \text{ tobacco acreage} = \frac{1940 \text{ tobacco acreage in MCD}}{1940 \text{ all land in farms in MCD}}
\]

\[
\% \text{ non-white operators} = \frac{\text{number of non-white operators in MCD in 1940}}{\text{number of farms in MCD in 1940}}
\]

\[
\% \text{ telephones} = \frac{\text{number of farms reporting telephones in MCD in 1940}}{\text{number of farms in MCD in 1940}}
\]

3. SAMPLING PLAN

In an actual survey we would employ a two stage sampling design with the minor civil divisions as the primary sampling units, and the master sample segments as the sub-sampling units. It is our desire to estimate a characteristic of the farm population in 1945, using a sampling scheme based on 1940 data. The census data for 1945 are then used to calculate variances so that the efficiency of the various stratifications may be measured. The sampling plan takes the following form.
a. Stratify the minor civil divisions in North Carolina into $K$ strata of
equal size.

b. From each stratum, select one MCD with probability proportionate to size,
where the measure of size is the 1940 number of farms.

c. Select the same proportion, $t$, from each stratum; that is, a constant
expected number of farms from whichever MCD is selected within a given
stratum.

d. Sub-sample master sample segments systematically from within the MCD
selected in step (c).

Illustration: Consider stratum number 100 in the stratification based on $\%$
non-white operators. This stratum is made up of five minor civil divisions of which
one is to be sampled with probability proportionate to size. If our predetermined
sampling rate is 5%; that is, $t = 1/20$, then we find the number of master sample
segments in our sub-sample in the following manner. The expected number of farms
in the five MCD's are:

<table>
<thead>
<tr>
<th>MCD</th>
<th>E.N.O.F.</th>
<th>CUMULATED</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>424</td>
<td>424</td>
</tr>
<tr>
<td>2</td>
<td>459</td>
<td>883</td>
</tr>
<tr>
<td>3</td>
<td>193</td>
<td>1076</td>
</tr>
<tr>
<td>4</td>
<td>150</td>
<td>1226</td>
</tr>
<tr>
<td>5</td>
<td>179</td>
<td>1405</td>
</tr>
</tbody>
</table>

draw a random number from 1 to 1405; let it be 1065. The MCD selected is number 3,
which covers the range from 883 to 1076. Since $P = \frac{P_{ij}}{P_j}$ is given as $\frac{193}{1405}$, and

$t = 1/20$, the number of master sample segments in the sample from this stratum is

1/ See footnote 1 page 4.

2/ In cases where the within strata variances are known or can be estimated with
reasonable accuracy, the size of the sample from each stratum can be gotten by
optimum allocation. As is shown by Neyman (4), the minimum variance is achieved
when the size of the sample from each particular stratum is taken in such a way that
it is proportional to the product of the population number of sub-sampling units in
the stratum with the standard deviation of that stratum. This would be a more
desirable method of deciding the size of sample from each stratum.
\[ m_{ij} = \frac{1}{20} \cdot \frac{1405}{193} M_{ij}, \]

where \( M_{ij} \) is the number of master sample segments in the \( i^{th} \) MCD of the \( j^{th} \) stratum.

We wish to estimate \( x \), the total of a characteristic in 1945, where

\[ x = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \sum_{\alpha=1}^{M_{ij}} x_{\alpha ij}, \]

and where \( x_{\alpha ij} \) is the total of a characteristic in the \( \alpha^{th} \) master sample segment of the \( i^{th} \) MCD in the \( j^{th} \) stratum. There are \( K \) strata with \( N_j \) minor civil divisions in the \( j^{th} \) stratum, and \( M_{ij} \) master sample segments in the \( i^{th} \) MCD of the \( j^{th} \) stratum. It should be pointed out that the number of strata, MCD's, and master sample segments do not change from 1940 to 1945.

To estimate \( x \), we use the ratio estimate

\[ (3.1) \quad \frac{x'}{r} = y \cdot \frac{x'}{y'}, \]

Here \( x' \) is an unbiased estimate of \( x \), and is given by the equation

\[ (3.2) \quad x' = \frac{1}{t} \sum_{i=1}^{K} \sum_{j=1}^{N_j} m_{ij} \bar{x}_{iij}, \]

where \( m_{ij} = \) number of master sample segments sampled from the \( i^{th} \) MCD of the \( j^{th} \) stratum,

\( \bar{x}_{iij} = \) sample mean per segment of a characteristic in the \( i^{th} \) MCD of the \( j^{th} \) stratum;

\( t = \) sampling rate.

The number of sampled segments is calculated by

\[ (3.3) \quad n_{ij} = \frac{tM_{ij}}{P}, \quad \text{where} \quad \frac{tM_{ij}}{P} \]

1/ Actually, the MCD's may change from one year to the next. Such was the case with the particular data which are used in this report. Necessary adjustments were made, however, before the data were used, so that comparable areas were maintained from the 1940 census to the 1945 census.
This probability is found as follows:

\[ P = \frac{P_{ij}}{P_j} = \frac{\text{number of farms in the } i^{th} \text{ MCD of the } j^{th} \text{ stratum in 1940}}{\text{number of farms in the } j^{th} \text{ stratum in 1940}} \]

The denominator of the ratio estimate, namely \( y' \), is the unbiased estimate of \( y \), where \( y \) is either the total number of farms in 1945 or the total of a certain characteristic in 1940, according as the ratio to estimated number of farms at the time of the survey or the ratio to the same characteristic at a previous census is used.

The approximate value of the variance of the ratio estimate is evaluated in the appendix, and is found to be

\[ \text{(3.4) } \text{Var} (x') = \frac{x^2}{2} \left\{ \frac{\text{Var} (x')}{x^2} + \frac{\text{Var} (y')}{y^2} - \frac{2 \text{Cov} (x'y')}{xy} \right\}. \]

The approximation to the square of the coefficient of variation is thus

\[ \text{(3.5) } (\text{C.V.})^2 = \frac{\text{Var} (x')}{x^2} \simeq \frac{\text{Var} (x')}{x^2} + \frac{\text{Var} (y')}{y^2} - \frac{2 \text{Cov} (x'y')}{xy}. \]

The variance of the unbiased estimate, \( x' \), is given by

\[ \text{(3.6) } \text{Var} (x') = K \sum_{j=1}^{N_j} \left\{ \sum_{i=1}^{P_{ij}} \frac{N_j P_{ij}}{M_{ij} - m_{ij}} \frac{(M_{ij} - m_{ij})^2}{(M_{ij} - 1) m_{ij}} + \sum_{i=1}^{N_j P_{ij}} \frac{M_{ij} P_{ij}}{m_{ij}} (\mu_{ij} - \mu_j)^2 \right\}, \]

where \( \mu_{ij} \) = true mean of a particular characteristic in the \( i^{th} \) MCD of the \( j^{th} \) stratum,

\( \mu_j \) = true mean of a particular characteristic in the \( j^{th} \) stratum.

The first term of (3.6) gives the between master sample segment within MCD contribution, and the second term the between MCD within stratum contribution to the overall variance. All the calculations and results in this report are based
on the between-MCD variance only. It is assumed that this term dominates the total variance. Using formula (3.5) it can be shown that the square of the between-MCD contribution to the coefficient of variation is

\[
(C.V.)^2_{1,1} = \frac{\sum_j \left( \frac{1}{\Sigma x_j} \sum_z R_{xz} Z_j \right)^2 - \sum_j x_j^2 \left( \Sigma y_j \right)^2 - \sum_j y_j^2}{\left( \Sigma x_j \right)^2 \left( \Sigma y_j \right)^2}
\]

The notation of formula (3.7) is defined below according to which ratio estimate is being used.

**Ratio to 1945 Farms**

\[
R_x = \frac{(1945 \text{ characteristic in p.s.u.})^2}{1940 \text{ farms in p.s.u.}}
\]

\[
R = \frac{(1945 \text{ farms in p.s.u.})^2}{1940 \text{ farms in p.s.u.}}
\]

\[
R_{xy} = \frac{1945 \text{ characteristic in p.s.u.}}{1940 \text{ farms in p.s.u.}} \times \frac{1945 \text{ farms in p.s.u.}}{1940 \text{ farms in p.s.u.}}
\]

**Ratio to 1940 Characteristic**

\[
R_x = \frac{(1940 \text{ characteristic in p.s.u.})^2}{1940 \text{ farms in p.s.u.}}
\]

\[
R = \frac{(1940 \text{ characteristic in p.s.u.})^2}{1940 \text{ farms in p.s.u.}}
\]

\[
R_{xy} = \frac{1945 \text{ characteristic in p.s.u.}}{1940 \text{ farms in p.s.u.}} \times \frac{1940 \text{ characteristic in p.s.u.}}{1940 \text{ farms in p.s.u.}}
\]

x = 1945 characteristic total in stratum

y = 1945 farms in stratum

z = 1940 farms in stratum

It should be pointed out, however, that Jube (5) found that the master sample segments as they are presently defined in North Carolina introduce a large within county variance for the linear unbiased estimate, for the ratio estimate of sample total of item to sample number of farms, and for the ratio estimate utilizing sample total of item to sample total of item at a previous date.
4. CALCULATIONS AND RESULTS

In considering a comparison of the estimates based on our sampling plan with those based on the sampling plan of Madow (1), several important points must be brought out. With the county as a primary sampling unit, Madow asserts that the between-primary sampling unit variance dominates the overall variance. We similarly assume that the between-MCD variance is the dominant term in the total variance, where the MCD is the primary sampling unit. It is also reasonable to assume that the between segment within county variance, which is represented by the scatter of master sample segments within the county, is at least as large as the between segment within MCD variance, which is represented by the scatter of segments within the MCD. This becomes evident when the size of the county is compared to the size of the MCD.

Now, if on the basis of our calculations we were to find that the between county variance is larger than the between MCD variance, then we could definitely conclude that the overall variance based on the county as a primary sampling unit is greater than the overall variance based on the MCD as a primary sampling unit. This can be illustrated algebraically as follows:

\[
\text{Var(TC)} = \text{Var(WC)} + \text{Var(BC)}
\]

\[
\text{Var(TM)} = \text{Var(WM)} + \text{Var(BM)}
\]

where \( \text{Var(TC)} \) is the total variance based on the county,

\( \text{Var(WC)} \) is the between segment within county variance,

\( \text{Var(BC)} \) is the between county variance,

\( \text{Var(TM)} \) is the total variance based on the MCD,

\( \text{Var(WM)} \) is the between segment within MCD variance,

\( \text{Var(BM)} \) is the between MCD variance.

1/ See footnote page 10
If \( \text{Var}(BC) > \text{Var}(BM) \) and \( \text{Var}(WC) \geq \text{Var}(WM) \) then \( \text{Var}(TC) > \text{Var}(TM) \).

In fact \( \text{Var}(TC) > \text{Var}(TM) \) if \( \text{Var}(BC) > \text{Var}(BM) \) and if \( \text{Var}(BC) - \text{Var}(BM) > |\text{Var}(WC) - \text{Var}(WM)| \).

The question still remains, however, to arrive at some final conclusions on the basis of the size of the variances. Can we say that 200 MCD's are better than 20 counties in yielding "best" overall estimates? An answer to this question involves the consideration of three important points: namely, gains due to stratification, cost, and size of the primary sampling unit.

For the purpose of calculating relative gains due to stratification, where the size of the primary sampling unit is the same, computations were made of the \((C.V.)^2\) in a sampling scheme where the MCD was used as a primary sampling unit in 20 strata. These figures appear in Table 1 along with the \((C.V.)^2\) of the items which are being estimated on the basis of 197 strata. Column (6) shows the percent gain of 197 strata over twenty strata, where the estimates are found by the ratio estimate method. Considerable gains are indicated for most items. However, Column (7) appears to be of more interest to us. Here the same comparison is made for the simple unbiased estimates, and greater gains of 197 strata over 20 strata are indicated. In investigating a similar situation, Cochran (3) points out that the gain in the use of the ratio estimate over the unbiased estimate decreases as the stratification is increased, where the stratification and the method of estimation are based on the size of unit. In fact, a point exists after which there is no further gain from the use of the ratio estimate over the unbiased estimate. Although our problem does not parallel Cochran's exactly, the results we get correspond to the statements which he makes.

A comparison of columns 3 and 5 with columns 4 and 6 indicates that a point exists somewhere between 20 strata and 197 strata after which the unbiased estimate is better than the ratio estimate. Increased stratification may therefore be more effective than the use of a complicated estimation method, "since if there are
Table 1. Comparison of Stratifications Based on Primary Sampling Units of Different Size

<table>
<thead>
<tr>
<th>P.S.U.</th>
<th>County 2/</th>
<th>MCD 3/</th>
<th>MCD 4/</th>
<th>% Gain of 197 MCD over 20 MCD 5/</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Strata</td>
<td>20</td>
<td>20</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td>Estimate Used</td>
<td>Unbiased : Ratio</td>
<td>Unbiased : Ratio</td>
<td>Unbiased : Ratio</td>
<td>Unbiased : Ratio</td>
</tr>
<tr>
<td>Item</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>No. of Tobacco Farms</td>
<td>.0048</td>
<td>.0047</td>
<td>.0098</td>
<td>.0085</td>
</tr>
<tr>
<td>Acres of Tobacco</td>
<td>.0058</td>
<td>.0061</td>
<td>.0127</td>
<td>.0125</td>
</tr>
<tr>
<td>Value of Lands and Buildings</td>
<td>.0027</td>
<td>.0020</td>
<td>.0096</td>
<td>.0070</td>
</tr>
<tr>
<td>No. of Cotton Farms</td>
<td>.0075</td>
<td>.0080</td>
<td>.0151</td>
<td>.0146</td>
</tr>
<tr>
<td>Acres of Cotton</td>
<td>.0182</td>
<td>.0184</td>
<td>.0324</td>
<td>.0263</td>
</tr>
<tr>
<td>No. of Non-white Operators</td>
<td>.0063</td>
<td>.0054</td>
<td>.0160</td>
<td>.0138</td>
</tr>
<tr>
<td>No. of Tenants</td>
<td>.0017</td>
<td>.0016</td>
<td>.0068</td>
<td>.0056</td>
</tr>
</tbody>
</table>
1/ The ratio to 1945 farms is used for the ratio estimate.

2/ The figures in columns (1) and (2) are those found in Madow (1). They are based on a sampling plan in which 20 strata are used, and in which the county is the primary sampling unit.

3/ The strata in columns (3) and (4) are the same 20 strata employed by Madow. The figures are based on a sampling plan in which the MCD is the primary sampling unit.

4/ The figures in columns (5) and (6) are based on the purely geographic stratification which is referred to in this report.

5/ Calculated as follows:

\[
\% \text{ gain} = \frac{1 - \left(\frac{C.V._{20}}{20}\right)^2}{\left(\frac{C.V._{197}}{197}\right)^2} \times 100, \text{ where } \frac{1}{20} = 0.05, \frac{1}{197} = 0.00515
\]

and where \(C.V._{20}\) represents the figures in either columns (3) or (4) and \(C.V._{197}\) represents the figures in either columns (5) or (6) according as the unbiased or ratio estimates are being compared.
enough strata, stratification will eliminate the effects of almost any kind of relation between \( y \) and \( x \)." Thus we conclude that if 197 strata are to be used, the simple unbiased estimate will be better than the ratio estimate in most cases. The gains which result in the use of either method would indicate that in preparing estimates of the farm population of North Carolina, if the MOD is the primary sampling unit, 197 strata will yield "better" estimates than 20 strata.

Again we point out that we define one estimate to be "better" than another only if the variance of the first is smaller than the variance of the second. For practical purposes, however, smaller variances will not necessarily be indicative of better estimates. Usually the objective in planning a survey is to get a minimum variance for a fixed cost. If the cost per schedule is fixed before the survey is run, the number of samples in the survey will be limited. In this case, samples from 197 strata may yield a variance which is larger than the variance obtained from a smaller number of strata. Furthermore, in an actual survey of the farm population of North Carolina it may be much more costly to sample an equal number of schedules from 197 strata than from 20 strata. Since the cost of a survey is an important factor in determining the extent of the survey, certain considerations other than those of a small variance must be made before determining a best sampling scheme. For instance, it may be found that for North Carolina a sampling scheme in which 100 strata are used will yield best estimates at a minimum cost. Such a plan remains to be investigated.

The efficiency of the county over the MOD as a primary sampling unit, where the number of strata is the same, can be measured by comparing columns (1) and (2) with columns (3) and (4) in Table 1. These figures seem to indicate that the county is a much more efficient unit than the MOD in a sampling scheme where 20 strata are used for preparing estimates of the farm population of North Carolina. However,

---

1/ Cochran (3), p. 128
2/ The problem of maximizing the amount of information obtainable from a given expenditure is discussed by Jessen (6).
where both the size of the primary sampling units and the number of strata differ, a comparison of the efficiency of the two different sampling schemes involves certain difficulties. The figures in columns (1) and (2) are based on a sampling plan in which the county is the primary sampling unit from 20 strata. These figures are to be compared with those which appear in columns (5) and (6), and which are based on a sampling plan in which the MOD is the primary sampling unit from 197 strata. By virtue of the arguments presented above, we know that some of the gain of the figures in columns (5) and (6) over those in columns (1) and (2) is due to stratification. In absence of calculations for the between-segment within primary sampling unit variances however, one hesitates to compare the efficiency of two sampling schemes in which the sizes of the primary sampling units differ. In this case it is not valid to say how much "better" one estimate is than the other on the basis of the between variances only. It is rather necessary to compare the total variances before any definite conclusion can be drawn. Further investigation of the between segment within primary sampling unit variances and of the costs of the individual sampling plans will be helpful in solving this problem.

The question of what type of stratification yields the "best" overall estimates still remains to be discussed. Results found in Table 2 indicate the following:

(a) In general, the best stratification to use for estimating any particular characteristic is a stratification by that characteristic. This point is evident in two instances in Table 2; namely stratification by \% tobacco acres and by \% non-white operators.

(b) A stratification based on a socio-economic characteristic, such as \% non-white operators or \% telephones, generally yields much poorer estimates of agricultural items, and only slightly, if at all, better estimates of its own characteristic than does a stratification based on tobacco acres or a geographic stratification in North Carolina. Where stratifications other than the purely geographic are used, the ratio
Table 2. Comparison of Four Different Types of Stratification for "Best" Overall Estimates

<table>
<thead>
<tr>
<th>Stratification Estimate Used</th>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purely Geographic: Tobacco Acres: Non-white Oper.: % Telephones</td>
<td>No. of Tobacco Farms:</td>
<td>0.000401: 0.000463: 0.000129: 0.000059: 0.000703: 0.000638: 0.000739: 0.000642</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acres of Tobacco</td>
<td>0.000645: 0.000722: 0.000299: 0.000325: 0.001029: 0.000990: 0.001021: 0.000963</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Value of Lands and Buildings</td>
<td>0.000585: 0.000567: 0.000625: 0.000453: 0.000723: 0.000554: 0.000642: 0.000478</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No. of Cotton Farms</td>
<td>0.000554: 0.000690: 0.000938: 0.000926: 0.001191: 0.001142: 0.001247: 0.001216</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acres of Cotton</td>
<td>0.001150: 0.001329: 0.002141: 0.002127: 0.002154: 0.002093: 0.002948: 0.002909</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No. of Non-white Operators</td>
<td>0.000807: 0.000778: 0.001321: 0.001125: 0.000505: 0.000283: 0.001228: 0.001060</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No. of Tenants</td>
<td>0.000421: 0.000499: 0.000450: 0.000388: 0.000544: 0.000458: 0.000545: 0.000458</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
estimate shows a gain over the unbiased estimate in general. This is explained by the fact that stratification on the basis of one characteristic may introduce heterogeneity among the other characteristics within the strata. In this case the heterogeneity appears to be taken care of more by the ratio estimate than by the unbiased estimate.

(c) We are concerned now with comparing the purely geographic with the \( \frac{\%}{\text{tobacco acres}} \) stratification. We note that in general, for the purely geographic stratification, the unbiased estimate is as good as the ratio to 1945 farms estimate, and usually slightly better, in the sense of smallest \( (\text{C.V.})^2 \). For the \( \frac{\%}{\text{tobacco acres}} \) stratification, the ratio estimate is worse for tobacco acres, much better for tobacco farms, and not very much better for the other items. Since the unbiased estimate is much easier to compute, it seems reasonable to use it if either the geographic or the \( \frac{\%}{\text{tobacco acres}} \) stratification is used.

(d) Hence we are led to decide upon the geographic versus the \( \frac{\%}{\text{tobacco acres}} \) stratifications on the basis of columns (1) and (3). The figures in these columns indicate that the geographic stratification is better except for tobacco items. We can of course obtain a lower \( (\text{C.V.})^2 \) by using the \( \frac{\%}{\text{tobacco acres}} \) stratification with the ratio estimate instead of the geographic stratification with the unbiased estimate for value of lands and buildings and for number of tenants, but it is doubtful that the slight reduction in the \( (\text{C.V.})^2 \) is worth the extra computation of using the ratio rather than the unbiased estimate.

(e) One interesting result in Table 2 was the smaller \( (\text{C.V.})^2 \) obtained in columns (3) and (4) for the number of tobacco farms than for acreage of tobacco, the item upon which the stratification was based. No immediate reason for this result is evident in the data.
The square of the coefficient of variation for value of lands and buildings and for number of non-white operators was also calculated by the ratio estimate, where the ratio was taken to the same characteristic in the 1940 census. Values obtained by this method for the purely geographic stratification indicate that "better" estimates may result from this estimation procedure than from any other. In both cases the resulting \( (O.V.)^2 \) were smaller than for the simple unbiased estimate and the ratio to 1945 farms estimate. However, the use of this procedure is usually limited by the lack of previous census data on the master sample segments, and also by the fact that units or items may be defined differently from one census to the next.

It should be pointed out that the conclusions stated above are based only on the data which have been examined for this report. Further investigation may show that some other stratification will be more desirable than the ones which were used here.

5. APPENDIX: FORMULAE & NOTATION

The derivations of the formulae for the expected values and the variances which appear in the preceding discussion are presented in the following pages.

Notation: Let

\[
\begin{align*}
x &= \text{total of a characteristic for the state of North Carolina in 1945,} \\
x_j &= \text{total of a characteristic in the } j\text{th stratum,} \\
x_{ij} &= \text{total of a characteristic in the } i\text{th MCD of the } j\text{th stratum,} \\
x_{\alpha ij} &= \text{total of a characteristic in the } \alpha\text{th master sample segment of the } i\text{th MCD of the } j\text{th stratum,}
\end{align*}
\]

\[1/ \text{Jebb (5) suggests a solution to this problem in the use of the "semi-ratio" estimate.}\]
where
\[ j = 1, 2, \ldots, K \]
\[ i = 1, 2, \ldots, N_j \]
\[ \alpha = 1, 2, \ldots, M_{ij} \]

There are \( K \) strata, the \( j \)th of which contains \( N_j \) MCD's, the \( i \)th of which contains \( M_{ij} \) master sample segments. Thus we can write

\[
x = \sum_{j=1}^{K} \left( \sum_{i=1}^{N_j} \left( \sum_{\alpha=1}^{M_{ij}} x_{ij} \right) \right)
\]

We also define

\( M_{ij} \) = number of master sample segments in the \( i \)th MCD of the \( j \)th stratum,
\( M_j \) = number of master sample segments in the \( j \)th stratum,
\( m_{ij} \) = number of master sample segments sampled from the \( i \)th MCD of the \( j \)th stratum,
\( P = \frac{P_{ij}}{P_j} \) = probability of selecting the \( i \)th MCD of the \( j \)th stratum with p.p.s.,
\( p \) = sampling rate,
\( \sigma^2 \) = variance of the \( i \)th MCD of the \( j \)th stratum,
\( \bar{x}_{ij} \) = sample mean of a particular characteristic in the \( i \)th MCD of the \( j \)th stratum,
\( \mu_{ij} \) = true mean of a particular characteristic in the \( i \)th MCD of the \( j \)th stratum,
\( \mu_j \) = true mean of a particular characteristic in the \( j \)th stratum.

A. Unbiased Estimate:

The total to be estimated is

\[
(A.1) \quad x = \sum_{j=1}^{K} \left( \sum_{i=1}^{N_j} \sum_{\alpha=1}^{M_{ij}} x_{ij} \right)
\]
The unbiased estimate of $x$ is,

$$x' = \frac{1}{t} \sum_{j=1}^{K} \sum_{i=1}^{N_j} m_{ij} \bar{x}_{ij}.$$  \hspace{1cm} (A.2) 

The sampling rate for the $i$th MOD in the $j$th stratum is

$$t_{ij} = \frac{P_j}{P_{ij}}.$$  \hspace{1cm} (A.3) 

where $P = \frac{P_{ij}}{P_j} = \frac{1940 \text{ farms in the } i\text{th MOD of the } j\text{th stratum}}{1940 \text{ farms in the } j\text{th stratum}}$

Applying (A.3) to the number of master sample segments, $M_{ij}$, in the selected MOD, we find that the number of m.s.s. in the sample is

$$M_{ij} = t_{ij} \frac{M_{ij} P_j}{P_{ij}}.$$  \hspace{1cm} (A.4)

Since the values of all the terms on the right of equation (A.4) are known, the number of master sample segments in the sample is exactly determined.

Proof that $x'$ is an unbiased estimate of $x$:

$$E(x') = E\left\{ \frac{1}{t} \sum_{j=1}^{K} \sum_{i=1}^{N_j} m_{ij} \bar{x}_{ij} \right\}$$

$$= \frac{1}{t} \sum_{j=1}^{K} \sum_{i=1}^{N_j} E(m_{ij} \bar{x}_{ij})$$

$$= \frac{1}{t} \sum_{j=1}^{K} \sum_{i=1}^{N_j} m_{ij} E\left( \frac{\bar{x}_{ij}}{m_{ij}} \right)$$

Applying (A.4) we have

$$E(x') = \sum_{j=1}^{K} \frac{P_j}{P} \sum_{i=1}^{N_j} \frac{M_{ij}}{P_{ij}} \frac{\bar{x}_{ij}}{m_{ij}}.$$  \hspace{1cm} (A.5)

But $E \bar{x}_{ij} = \sum_{i=1}^{N_j} P_{ij} / i_{ij} P_j$.
Hence \( (A.7) \ E(x^t) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} M_{ij}/\mu_{ij} \).

Now \( /\mu_{ij} = \frac{M_{ij}}{\sum_{i=1}^{\infty} M_{ij}} \).

Therefore \( (A.7) \) becomes

\[
(A.8) \ E(x^t) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \frac{M_{ij}}{\sum_{i=1}^{\infty} M_{ij}} \alpha_{ij} = x
\]

The error in the estimate is given by

\[
(A.9) \ n_{ij} \bar{x}_{ij}/n - M_j /\mu_j
\]

where \( \sum_{j=1}^{K} M_j /\mu_j \) is the true population total.

Applying \( (A.4) \) to \( (A.9) \) we have the error expressed as

\[
(A.10) \ M_{ij} P_{ij} \bar{x}_{ij}/P_{ij} - M_j /\mu_j
\]

which may be written as

\[
(A.11) \ \left\{ \frac{M_{ij} P_{ij}}{P_{ij}} (\bar{x}_{ij} - \mu_{ij}) \right\} + \left\{ \frac{M_{ij} P_{ij}}{P_{ij}} \mu_{ij} - M_j /\mu_j \right\}.
\]

Thus the variance of the unbiased estimate is

\[
(A.12) \ \mathbb{E} \left[ \left\{ \frac{M_{ij} P_{ij}}{P_{ij}} (\bar{x}_{ij} - \mu_{ij}) \right\} + \left\{ \frac{M_{ij} P_{ij}}{P_{ij}} \mu_{ij} - M_j /\mu_j \right\} \right]^2,
\]

which is equal to

\[
(A.13) \ \mathbb{E} \left[ \frac{M_{ij} P_{ij}}{P_{ij}} (\bar{x}_{ij} - \mu_{ij}) \right]^2 + \mathbb{E} \left[ \frac{M_{ij} P_{ij}}{P_{ij}} \mu_{ij} - M_j /\mu_j \right]^2.
\]
The cross-product term drops out because the selection of master-sample segments within one MCD is independent of the selection within any other MCD.

In evaluating the terms of (A.13), we weight each by a factor of $P_{ij}/P_j$, which represents the probability of selection. Thus we can write the variance of the unbiased estimate as

\[
(A.14) \ \text{Var} \ (x') = \sum_{j=1}^{K} \left( \frac{N_j P_{ij}}{P_j} \right) \frac{M_{ij}^2 P_j}{P_{ij}^2} \left( 1 - \frac{m_{ij} - 1}{M_{ij} - 1} \right) \frac{\sigma_{ij}^2}{m_{ij}} + \sum_{j=1}^{K} \left( \frac{M_{ij} P_j}{P_{ij}} \left( \frac{\mu_{ij} - N_j}{\mu_{ij} - M_{ij}} \right) \right)^2 ,
\]

or

\[
(A.15) \ \text{Var} \ (x') = \sum_{j=1}^{K} \left( \frac{N_j M_{ij}^2 P_j}{P_{ij}^2} \left( \frac{M_{ij} - m_{ij}}{M_{ij} - 1} \right) \frac{\sigma_{ij}^2}{m_{ij}} \right) + \sum_{j=1}^{K} \left( \frac{M_{ij} P_j}{P_{ij}} \left( \frac{\mu_{ij} - N_j}{\mu_{ij} - M_{ij}} \right) \right)^2 ,
\]

where

\[
\sigma_{ij}^2 = \frac{\sum_{\alpha=1}^{M_{ij}} (x_{\alpha ij} - \mu_{ij})^2}{M_{ij} - 1}.
\]

The first term in equation (A.15) gives the within-MCD variance, while the second term gives the between-MCD variance. The calculations which appear in this paper are based only on the between-MCD contribution to the overall variance. It is assumed that this term dominates the total variance.

For ease of calculation the second term of (A.15), namely the between-MCD variance, is reduced in the following manner.

\[
(A.16) \ \sum_{j=1}^{K} \left( \frac{N_j M_{ij}^2 P_j}{P_{ij}^2} \left( \frac{M_{ij} - m_{ij}}{M_{ij} - 1} \right) \frac{\sigma_{ij}^2}{m_{ij}} \right) = \sum_{j=1}^{K} \left( \frac{M_{ij}^2 P_j}{P_{ij}^2} \left( \frac{\mu_{ij} - N_j}{\mu_{ij} - M_{ij}} \right) \right)^2 - 2 \sum_{j=1}^{K} M_{ij} \mu_{ij}^2 \approx \sum_{j=1}^{K} \frac{M_{ij}^2}{P_{ij}} \mu_{ij}^2 \frac{N_j}{M_{ij} - 1} + \frac{K M_{ij}^2}{P_{ij}} \sum_{j=1}^{K} \frac{N_j}{M_{ij} - 1} \mu_{ij} P_{ij} .
\]
Note that

\[ \mu_{ij} = \frac{x_{ij}}{n_{ij}} \quad \text{and} \quad \mu_j = \frac{x_j}{n_j} \]

Replacing these values in (A.16), we get

\[ (A.17) \sum_{j=1}^{K} P_j \sum_{i=1}^{N_j} x_{ij}^2 - 2 \sum_{j=1}^{K} x_j \sum_{i=1}^{N_j} x_{ij} + \sum_{j=1}^{K} \frac{x_j^2}{n_j} \sum_{i=1}^{N_j} P_{ij}, \]

which reduces further to

\[ (A.18) \sum_{j=1}^{K} P_j \frac{N_j x_{ij}^2}{n_j} - \sum_{j=1}^{K} x_j^2, \]

because \[ \sum_{j=1}^{N_j} x_{ij} = x_j \quad \text{and} \quad \sum_{i=1}^{N_j} P_{ij} = P_j \]

Thus the calculations for the variance of the unbiased estimate are based on formula (A.18), and give only the between MCD variance.

B. Ratio Estimate:

To estimate \( x \), the total of a characteristic in 1945, we use the ratio estimate,

\[ (B.1) x' = \frac{x'}{y'} \quad \text{where} \quad x' \quad \text{is an unbiased estimate of} \quad x \quad \text{and is given by} \]

formula (A.2), and where \( y' \) is similarly the unbiased estimate of \( y \). \( y \) is defined either as the total number of farms in 1945, or as the total of the characteristic in 1940, according as the ratio to the estimated number of farms at the time of the survey or the ratio to the same characteristic at a previous census is used.
The variance of $x_1^t$ is approximated in the following manner. Let

$$(B.2) \Delta x = x_1^t - x \text{ and } \Delta y = y_1^t - y.$$  

Then $(B.3)$ $E(\Delta x)^2 = 0$, $E(\Delta y) = 0$, 

$$E(\Delta x)^2 = \sigma_x^2, \ E(\Delta y)^2 = \sigma_y^2,$$

and $E(\Delta x)(\Delta y) = \sigma_{xy}^2$.

Applying $(B.2)$ to $(B.1)$ we can write the ratio estimate

$$(B.4) \ x_1^t = y \cdot \left\{ \frac{x + \Delta x}{y + \Delta y} \right\}, \text{ which is equivalent to}$$

$$(B.5) \ x_1^t = x \left\{ 1 + \frac{\Delta x}{x} \right\} \left\{ 1 + \frac{\Delta y}{y} \right\}^{-1}.$$

Upon expanding the last term in $(B.5)$, we get

$$(B.6) \ x_1^t = x \left\{ 1 + \frac{\Delta x}{x} \right\} \left\{ 1 - \frac{\Delta y}{y} + \frac{(\Delta y)^2}{y^2} - \ldots \right\}$$

which is approximately given by the relationship

$$(B.7) \ x_1^t \approx x \left\{ 1 + \frac{\Delta x}{x} - \frac{\Delta y}{y} \right\}.$$

Taking the expected values of both sides of $(B.7)$ we find that

$$(B.8) \ E x_1^t \approx x.$$

Combining $(B.7)$ and $(B.8)$ we get

$$(B.9) \ x_1^t \approx E x_1^t \approx x \left\{ \frac{\Delta x}{x} - \frac{\Delta y}{y} \right\}.$$

The variance of the ratio estimate is found by taking the expectation of the square of equation $(B.9)$. This yields

$$(B.10) \ E(x_1^t - E x_1^t)^2 \approx x^2 \ E \left\{ \frac{\Delta x}{x} - \frac{\Delta y}{y} \right\}^2.$$
Applying the expression in (B.3) we arrive at the approximation to the variance of $x_i'$. Thus

$$E(x_i' - E x_i')^2 \approx x_i^2 \left( \frac{\sigma^2}{x} + \frac{\sigma^2}{y} - \frac{2\sigma x_i y_i'}{xy} \right).$$

The approximation to the square of the coefficient of variation follows from (B.11), and is written

$$(B.12) \quad (C.V.)^2 = \frac{\text{var}(x_i')}{x^2} = \frac{\sigma^2}{x^2} + \frac{\sigma^2}{y^2} - \frac{2\sigma x_i y_i'}{xy}.$$ 

C. Covariance of the Two Unbiased Estimates

The covariance of $x_i'$ and $y_i'$ is found in the same manner as the variance of the unbiased estimate. Here the between-MCD covariance is given by:

$$(C.1) \quad \sum_{j=1}^{K} \sum_{i=1}^{N_j} \frac{P_{ij}}{P_j} \left( \frac{M_{ij} \mu_{ijx} - M_{ij} \mu_{ijx}}{P_{ij}} \right) \left( \frac{M_{ij} \mu_{ijy} - M_{ij} \mu_{ijy}}{P_{ij}} \right),$$

where $\mu_{ijx}$ and $\mu_{ijy}$ are population means of the $x$ variables, and $\mu_{ijy}$ and $\mu_{ijy}$ are population means of the $y$ variables.

After manipulation similar to those used in section (A) of this appendix, (C.1) reduces to

$$(C.2) \quad \sum_{j=1}^{K} \frac{P_j}{P} \sum_{i=1}^{N_j} \frac{\bar{x}_{ij} \bar{y}_{ij}}{P_{ij}} - \sum_{j=1}^{K} \frac{P_j}{P} \bar{x}_j \bar{y}_j.$$
REFERENCES


