

INVESTIGATION OF AREA SUB-STRATIFICATION AND THE EFFICIENCY
OF THE MINOR CIVIL DIVISION AS A PRIMARY SAMPLING
UNIT IN ESTIMATING AGRICULTURAL CHARACTERISTICS
IN NORTH CAROLINA

By

Joe Nelson Boyd

*Institute of Statistics
Memo. Series #31
For limited distribution*

TABLE OF CONTENTS

	<u>Page</u>
Introduction	1
Area Sub-stratification	3
Investigation of the Efficiency of the Minor Civil Division (M.C.D.) as a Primary Sampling Unit and Master Sample Segments as the Sub-sampling Units	16
Method of Stratification	16
Basic Sampling Plan	17
Calculations and Results	17
Unbiased Estimate	17
Ratio Estimate	18
Regression Estimate	19
Appendix	28
References	36

LIST OF TABLES AND MAPS

	<u>Page</u>
Table I. Comparison of the Coefficients of Variation and the Relative Efficiencies of Different Methods Using 20 Strata, Where One County is Selected with P.P.S. from Each Stratum	12
Table II. Bias of the Estimate in Per Cent for Area Sub-stratification with 20 Strata	13
Table III. Comparison of the Coefficients of Variation and the Relative Efficiencies of Different Methods Using 10 Strata Where One County is Selected with P.P.S. from Each Stratum	15
Table IV. Comparison of the Between Primary Sampling Unit Component of the (Coefficient of Variation) ² for Approximately 200 Strata of M.C.D.'s Using Different Methods of Estimation	21
Table V. Bias of the Ratio Estimate in Per Cent	22
Table VI. Comparison of the Between Primary Sampling Unit Component of the (Coefficient of Variation) ² for Stratification by Current Census Information with that for Past Census Information Using the Unbiased Estimate	23
Table VII. Comparison of the (Coefficient of Variation) ² of 20 Strata When the County is the Primary Sampling Unit and 20 and 197 Strata Where the M.C.D. is the Primary Sampling Unit	25
Map I. Types of Farming Areas in North Carolina with County Boundaries Preserved	4
Map II. Basic 10 and 20 Strata by County	5

INTRODUCTION

The work reported in this thesis is a continuation of a study reported by Lillian H. Madow [1] .

Mrs. Madow used a two stage sampling plan where the county was the primary sampling unit and the Master Sample segments were the secondary sampling units. ^{1/} If a simple unbiased estimate were used, it was determined that 20 counties were not sufficient to estimate most items with an accuracy of a 5 per cent coefficient of variation and even samples of 40 counties would not often yield estimates within a 5 per cent coefficient of variation.

As cost considerations usually limit most surveys to 20 counties or one-fifth the total number in North Carolina, Mrs. Madow proceeded to investigate the accuracy of estimates other than the simple unbiased estimate.

These other estimates were

- (1) ratio to estimated number of farms at time of survey
- (2) ratio to same characteristic at a previous census
- (3) regression estimate.

Little improvement was found by using the first estimate, but it was found that 20 counties were adequate for most items when the second was used. The third showed only slight gains over the second. The chief objection to these estimates is that they are more laborious to compute and in estimates (2) and (3) comparable data may not be available at an earlier Census date.

Since cost considerations usually limit most surveys to 20 counties and since estimates (2) and (3) require information which may not be

^{1/} Refer to King and Jessen [2] .

available, how can estimates based on 20 counties be improved? The research on this subject in this thesis was divided into two parts.

I. Investigation of the use of area sub-stratification where the Minor Civil Division (MCD) is the unit of sub-stratification.

II. Investigation of the efficiency of the MCD as a primary sampling unit and Master Sample segments as the sub-sampling units rather than the county as a primary unit and Master Sample segments as sub-sampling units as used previously.

In connection with II above, the following comparisons were made:

1. The ratio to 1945 number of farms estimate with an unbiased estimate and the bias of the ratio estimate.
2. A regression estimate with a ratio to 1945 number of farms estimate.
3. A purely geographic sub-stratification with four sub-stratifications by socio-economic characteristics.
4. A sub-stratification by current census information with that of past census information.
5. A stratification of MCD's in 200 strata with that of 20 strata where the MCD is the primary sampling unit.

The sub-stratification items used in these investigations were:

1. days worked off farms in 1940
2. value of products sold in 1940
3. average miles from all weather roads in 1945
4. farms operated by tenants in 1945.

The items to be estimated were:

1. tobacco acreage in 1945
2. cotton acreage in 1945
3. days worked off farms in 1945.

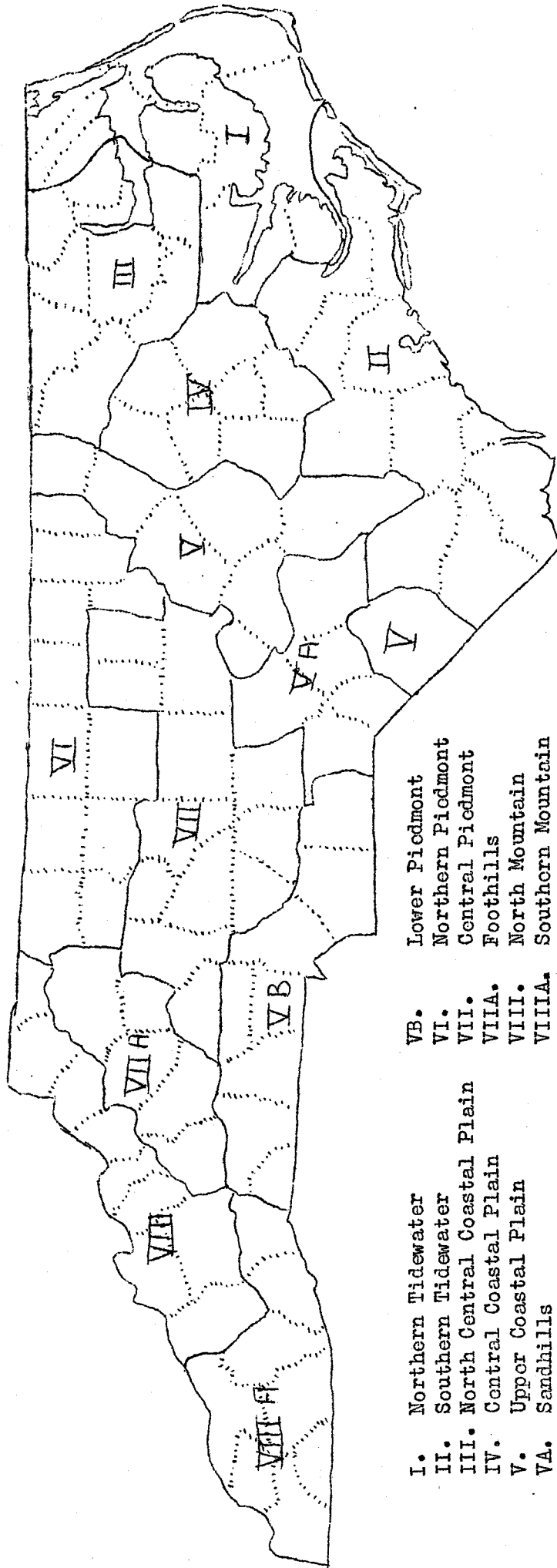
AREA SUB-STRATIFICATION

The purpose of area sub-stratification is to adjust the sample taken from the primary sampling unit (a county, in this part of the study) in such a way so as to make it representative of the stratum from which it was selected rather than representative of the primary sampling unit.

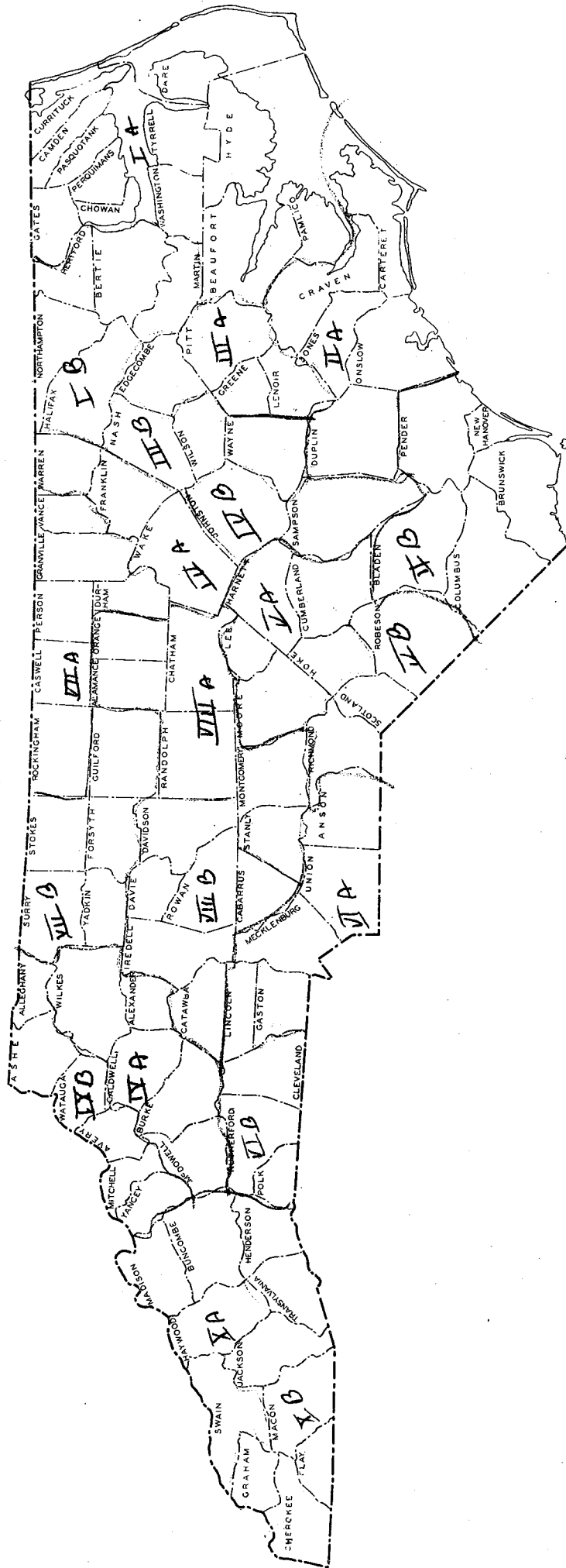
In order to use area sub-stratification, the population must be capable of being divided into primary sampling units which in turn must be capable of being divided into sub-areas. Also these sub-areas must be accompanied by certain Census information, such as on total farms in 1940 and on sub-stratification items. In this design the state of North Carolina is the population, the county is the primary sampling unit and the MCD (township) is the sub-area. The MCD will be used only to determine the sub-strata within the primary geographic stratum and the primary sampling unit, as Master Sample segments will be the sub-sampling units within the sub-strata of the primary sampling units.

For purposes of comparison the same 10 and 20 major geographic strata were used as in [1] (see Maps 1 and 2). The design assumes that one county is selected from each primary stratum with probability proportionate to the 1940 number of farms (p.p.s.). Each MCD within a given primary stratum was allocated to one of two sub-strata on the basis of one of the stratification items listed previously. These two sub-strata within the primary stratum contained approximately the same number of farms listed according to the 1940 Census. Hence each MCD in a given county was also allocated to one of the two sub-strata on the basis of the figures obtained to divide the primary stratum into two sub-strata, but the two sub-strata within a given county would not be expected to have the same number of farms. Hence this design would have a different sampling rate for the two

Map 1. Types of Farming Areas in North Carolina



Map 2. Basic Strata



sub-strata within each county selected. Master Sample segments were the sub-sampling units within each sub-stratum of MCD's within the county selected.

For purposes of illustration the following example is given. We will assume that the first stratification item, "Days Worked Off Farms in 1940," was used. The MCD's in primary stratum IXB were put in an array by the average of this item for each MCD and split into two sub-strata with approximately equal number of farms in each sub-stratum. Let us assume that a primary stratum IXB has 3 counties, Alleghany (1), Avery (2) and Mitchell (3). The pertinent data follow.

Primary Stratum IXB

County	MCD within county	1940 number of farms	Cumulated 1940 number of farms	Average number of days worked off farms in 1940
1	1	212	212	1.5
1	6	265	477	7.3
1	2	117	594	9.4
1	4	347	941	10.6
2	1	225	1166	15.4
1	5	227	1393	16.8
1	3	384	1777	19.4
3	3	203	1980	21.4
3	7	101	2081	29.2
3	8	94	2175	29.7
2	4	341	2516	43.0
2	5	269	2805	53.9
3	9	150	2955	54.1
2	6	170	3125	55.4
1	7	138	3263	56.8
3	5	472	3735	62.6
3	6	288	4023	63.4
3	2	171	4194	65.0
2	3	170	4364	67.5
2	2	176	4540	87.7
2	7	213	4753	114.5
3	4	102	4855	116.6
3	1	286	5141	126.8

It can be seen that the figure for average number of days worked off farms necessary to divide the primary stratum into two sub-strata with approximately equal number of farms is somewhere between 43 and 54. It should be noted that this figure will vary from primary stratum to primary stratum for each of the four stratification items used. One county is now selected from this primary stratum with probability proportionate to the 1940 number of farms (p.p.s.) and this is done by simply cumulating the number of farms by county and selecting a random number between 1 and 5141, the latter being the total number of farms for this particular primary stratum. This technique is shown below.

County	1940 number of farms	Cumulated 1940 number of farms
1	1690	1690
2	1584	3274
3	1867	5141

If 1807 is the random number selected, county 2 is the sample county. This county has 36% of its farms in sub-stratum I and 64% of its farms in sub-stratum II. Let the within stratum sampling rate be l/t , and the within county sampling rate be l/c , where c/t is the ratio of the number of farms in the sample county and the number of farms in the primary stratum from which the county was selected. Thus the sampling rate in Sub-stratum I will be $25/18c$ and the sampling rate in Sub-stratum II will be $25/32c$ and therefore the sampling rate within the county is

$$\frac{64}{100} \frac{25}{32c} = \frac{36}{100} \frac{25}{18c} + \frac{25}{18c}$$

It can be seen that two sampling rates are used within each county. The actual number of sub-strata used within each primary stratum can be greater than two provided each primary unit has an adequate representation of each sub-stratum. In this thesis, only two sub-strata have been used.

The following subscript notation will be used in the remainder of this thesis:

<u>Subscript</u>	<u>Subdivision</u>	<u>Number of subdivision in</u>	
		<u>population</u>	<u>sample</u>
α	α^{th} stratum	G	g
αi	i^{th} county in α^{th} stratum	H_{α}	1
αij	j^{th} substratum in (αi) county	2	2
αijs	s^{th} Master Sample unit in (αij) substratum	$M_{\alpha ij}$	$m_{\alpha ij}$

X will refer to the characteristic being estimated and X^i to the estimate of X . P will refer to the total number of farms in 1940. For example $X_{\alpha ij}$ is the total of a particular characteristic being estimated in the j^{th} substratum of the i^{th} county in the α^{th} stratum. If it is desired to indicate the total summed over an inner subscript, we shall indicate this by a dot. For example, $X_{\alpha \cdot j}$ is the sum over all counties in the j^{th} sub-stratum of the α^{th} stratum.

The population total to be estimated can be expressed as

$$(1) \quad X = \sum_{\alpha=1}^G \sum_{i=1}^{H_{\alpha}} \sum_{j=1}^2 \sum_{s=1}^{M_{\alpha ij}} X_{\alpha ijs}$$

A similar expression for this estimate of X is

$$(2) \quad X^i = \sum_{\alpha=1}^G \frac{1}{t_{\alpha}} \sum_{j=1}^2 \sum_{s=1}^{m_{\alpha ij}} X'_{\alpha ijs}$$

where t_{α} is the sampling rate within the α^{th} stratum.

t_{α} can be expressed as

$$(3) \quad t_{\alpha} = \frac{m_{\alpha ij} P_{\alpha ij}}{M_{\alpha ij} P_{\alpha \cdot j}}$$

$m_{\alpha ij}$ can be found from (3) to be

$$(4) \quad m_{\alpha ij} = M_{\alpha ij} t_{\alpha} P_{\alpha \cdot j} / P_{\alpha ij}$$

The expected value of X^i is

$$(5) \quad E X^i = \sum_{\alpha} \sum_i \sum_j \sum_s \frac{1}{t_{\alpha}} \frac{P_{\alpha i}}{P_{\alpha}} \frac{m_{\alpha ij}}{M_{\alpha ij}} X_{\alpha ijs}$$

Substituting $t_{\alpha} = m_{\alpha ij} P_{\alpha ij} / M_{\alpha ij} P_{\alpha \cdot j}$ into (5), we obtain

$$E X^i = \sum_{\alpha} \sum_i \sum_j \sum_s \frac{P_{\alpha \cdot j}}{P_{\alpha ij}} \frac{P_{\alpha i}}{P_{\alpha}} X_{\alpha ijs} \text{ which can be expressed, by}$$

multiplying and dividing by P_{α} as

$$E X^i = \sum_{\alpha} P_{\alpha} \sum_i \sum_j \frac{P_{\alpha i}}{P_{\alpha}} \frac{P_{\alpha \cdot j}}{P_{\alpha}} \frac{X_{\alpha ij}}{P_{\alpha ij}}$$

If we denote the double summation by $R_{\alpha}(A)$, the average of the adjusted ratios within the α^{th} stratum, we obtain

$$E X^i = \sum_{\alpha} P_{\alpha} R_{\alpha}(A) \quad \text{where } P_{\alpha} = \sum_i P_{\alpha i} = \sum_j P_{\alpha \cdot j}$$

$$R_{\alpha}(A) = \sum_i \frac{P_{\alpha i}}{P_{\alpha}} R_{\alpha i}(A) \quad \text{where } R_{\alpha i}(A) = \sum_j \frac{P_{\alpha \cdot j}}{P_{\alpha}} R_{\alpha ij}$$

$$\text{where } R_{\alpha ij} = \sum_s X_{\alpha ijs} / P_{\alpha ij} = \frac{X_{\alpha ij}}{P_{\alpha ij}}$$

Since $X = \sum_{\alpha} X_{\alpha} = \sum P_{\alpha} R_{\alpha}$ and $E X' = \sum P_{\alpha} (R_{\alpha(A)})$.

$$(6) \quad E X' = X + \sum_{\alpha} P_{\alpha} [R_{\alpha(A)} - R_{\alpha}] .$$

Hence the estimate X' is biased. This bias is the sum of the biases over all the primary strata. See reference [3] for a more complete discussion of the preceding theory.

The mean square error for area sub-stratification can be expressed as ^{1/}

$$(7) \quad \sigma_{X'}^2 = \sum_{\alpha} \sum_i \sum_j P_{\alpha \cdot j}^2 \frac{P_{\alpha i} (M_{\alpha ij} - m_{\alpha ij})}{P_{\alpha} (M_{\alpha ij} - 1)} \frac{\sigma_{\alpha ij}^2}{m_{\alpha ij} \bar{P}_{\alpha ij}^2} \\ + \sum_{\alpha} P_{\alpha}^2 \sum_i \frac{P_{\alpha i}}{P_{\alpha}} \left(R_{\alpha i(A)} - R_{\alpha(A)} \right)^2 + \left[\sum P_{\alpha} (R_{\alpha(A)} - R_{\alpha}) \right]^2$$

where

$$\sigma_{\alpha ij}^2 = \sum_s \left(X_{\alpha ijs} - \bar{X}_{\alpha ij} \right)^2 / M_{\alpha ij} \text{ and is the variance between Master}$$

Sample segments within the sub-strata within the county selected. Also

$$\bar{P}_{\alpha ij} = P_{\alpha ij} / M_{\alpha ij} .$$

Thus the mean square error is composed of three components (from left to right): (i) the between Master Sample segments component, (ii) the between county within primary strata component and (iii) the bias component. Previous studies have indicated that (ii) is the largest component with the importance of (i) depending on the sampling rate used. The bias component (iii) was expected to be negligible, but it turned out to be

^{1/} This is the same formula given in [3] with minor changes in notation.

relatively large for some items studied in this thesis.

In the discussion which follows, coefficients of variation for three methods of estimation with no sub-stratification and for four methods of stratification with area sub-stratification will be compared. Table 1 presents these seven coefficients of variation and the accompanying relative efficiencies for three items when the state was divided into 20 strata. The results for no sub-stratification were previously reported in [1]. For estimating "Days Worked Off Farms" all four methods for sub-stratification were less efficient than the last two with no sub-stratification. Strangely enough the smallest relative efficiency of the four sub-stratification items was "Days Worked Off Farms", the socio-economic characteristic to be estimated. It is believed that the reason for this is the shift from farm to city during the years 1940 to 1945 when the war industries were in full swing. For "Tobacco Acreage" only the second method with sub-stratification was more efficient than the unbiased estimate. The small relative efficiency for the fourth sub-stratification method is not in accordance with the opinion that per cent tenancy and tobacco acreage are highly correlated. For "Cotton Acreage" the first and fourth methods with sub-stratification were slightly more efficient than the unbiased estimate.

The bias component contributed a considerable amount to the mean square error in many cases as shown in Table 2. It is believed that the reason for the large bias in the case of "Days Worked Off Farms" is the shift of farm labor to factory labor during the war years. The large biases in estimating "Tobacco Acreage" and "Cotton Acreage", when the last two sub-stratification items were used, were probably attributable to a fairly consistent positive correlation between P_{cij}/P_{ci} (the ratio of the number of

Table 1. Comparison of the Coefficients of Variation and the Relative Efficiencies ^{1/} of Different Methods Using 20 Strata, Where One County is Selected with p.p.s. from Each Stratum.

Type of estimate	: Days worked		: Tobacco		: Cotton	
	: off farm		: acreage		: acreage	
	: in 1945		: in 1945		: in 1945	
	:C.V.:	:R.E. in %:	:C.V.:	:R.E. in %:	:C.V.:	:R.E. in %:
<u>No area sub-stratification</u>	:	:	:	:	:	:
Unbiased	:.152:	100	:.086:	100	:.135:	100
Ratio to 1945 number of farms	:.143:	113	:.088:	95	:.136:	99
Ratio to 1940 characteristic	:.114:	178	:.020:	1450	:.033:	1655
	:	:	:	:	:	:
<u>Area sub-stratification</u>	:	:	:	:	:	:
Number of days worked off farm in 1940	:.160:	91	:.077:	97	:.130:	108
Value of products sold in 1940	:.159:	91	:.073:	107	:.136:	98
Average miles from all weather roads in 1945	:.151:	102	:.081:	89	:.146:	86
Farms operated by tenants in 1945	:.146:	108	:.092:	68	:.133:	103

^{1/}

Relative efficiency (R.E.) of two different methods, according to Cochran [4], is defined as the inverse ratio of their variances. For example under "Days Worked Off Farms in 1945" the relative efficiency in per cent of the ratio to 1945 number of farms estimate to the unbiased estimate is

$$\frac{(C.V_2)^2}{(C.V_1)^2} \cdot 100 = \frac{\frac{V_2}{X^2}}{\frac{V_1}{X^2}} \cdot 100 = \frac{(.152)^2}{(.143)^2} \cdot 100 = 113\%$$

Table 2. Bias of the Estimate in Per Cent of 1945 Characteristic for Area Sub-stratification with 20 Strata.

Sub-stratification item:	1945 Estimate		
	Days worked : : off farms	Tobacco acreage	Cotton acreage
Number of days worked off farms in 1940	9.48	1.77	.19
Value of products sold in 1940	3.20	.45	.86
Average miles from all weather roads in 1945	4.63	3.23	2.24
Farms operated by tenants in 1945	5.84	4.18	3.54

farms in the sub-stratum to the number in the county), and R_{cij} (the mean characteristic per farm for the sub-stratum) within each sub-stratum. Hansen and Hurwitz point out that if there were no correlation, R_{α} would equal $R_{\alpha(A)}$; hence, there would be no bias [3] .

Table 3 shows a comparison of area sub-stratification with the unbiased estimate when 10 strata are used. For "Days Worked Off Farms" only the second sub-stratification method was less efficient than the unbiased estimate. For "Tobacco Acreage" only the third sub-stratification method was as efficient as the unbiased estimate. The fourth method again had a very low relative efficiency. For "Cotton Acreage" the first and fourth methods were more efficient than the unbiased estimate.

In the case of both 10 and 20 strata there is no evidence that any of the 4 sub-stratification methods used in this investigation is consistently more efficient than the unbiased estimate. Hence it would appear inadvisable to recommend area sub-stratification for the following two reasons:

(i) The estimate using area sub-stratification is generally biased, often by a large amount. Further the magnitude of this bias can not be estimated in an actual survey.

(ii). There is a certain amount of extra computing in drawing a sample when area sub-stratification is used.

Table 3. Comparison of the Coefficients of Variation and the Relative Efficiencies of Different Designs Using 10 Strata Where One County is Selected with p.p.s. from each Stratum.

Type of estimate	: Days worked		: Tobacco		: Cotton	
	: off farm		: acreage		: acreage	
	: in 1945		: in 1945		: in 1945	
	:C.V.:	R.E. in %	:C.V.:	R.E. in %	:C.V.:	R.E. in %
<u>No area sub-stratification</u>	:	:	:	:	:	:
Unbiased	:.238:	100	:.122:	100	:.214:	100
<u>Area sub-stratification</u>	:	:	:	:	:	:
Number of days worked off farm in 1940	:.232:	105	:.129:	90	:.212:	102
Value of products sold in 1940	:.245:	94	:.133:	84	:.219:	95
Average miles from all weather roads in 1945	:.232:	105	:.122:	100	:.227:	89
Farms operated by tenants in 1945	:.231:	106	:.152:	64	:.198:	117
	:	:	:	:	:	:
	:	:	:	:	:	:

INVESTIGATION OF THE EFFICIENCY OF THE MINOR CIVIL DIVISION (M.C.D.)
AS A PRIMARY SAMPLING UNIT WITH MASTER SAMPLE SEGMENTS AS THE
SUB-SAMPLING UNITS.

Methods of Stratification

The state of North Carolina was first divided into 12 major geographic strata according to type of farming areas. These 12 strata are shown in Map 1. It was possible to improve upon these original 12 strata by following M.C.D. boundaries rather than county boundaries and then using the M.C.D. as the primary sampling unit.

For the purely geographic stratification, contiguous M.C.D.'s were combined within each major geographic stratification in such a way that each stratum contained approximately 5 M.C.D.'s and 1400 farms. This resulted in obtaining approximately 200 strata within the 12 major geographic strata. ^{1/} For the other four types of stratification the M.C.D.'s within each of the 12 major geographic strata were put into an array by the magnitude of the type of stratification and combined in such a manner that each stratum contained approximately 1400 farms. It was believed that the size of sample with one M.C.D. drawn from each of the 200-odd strata would be approximately the same as with one county drawn from each of 20 strata (sub-sampling rate assumed to be equal). Dare and Swain counties were omitted from the population leaving 98 counties and 941 M.C.D.'s. ^{2/}

^{1/}

It was not possible to obtain exactly the same number of strata for each of the several types of stratification used in this investigation. This number varied from 194 to 197, but will be referred to as approximately 200 in the text.

^{2/}

If an M.C.D. had less than 100 farms in 1940 it was combined with an adjacent M.C.D. so that each of the 941 combined M.C.D.'s had at least 100 farms in 1940. There were a total of 1018 original M.C.D.'s in 1945.

Basic Sampling Plan

In an actual survey, from each of the approximately 200 strata one M.C.D. would be selected with probability proportional to the 1940 number of farms, and Master Sample segments then chosen systematically from each of the selected M.C.D.'s.

Calculations and Results

The total to be estimated can be expressed as

$$(8) \quad X = \sum_{\alpha=1}^G \sum_{i=1}^{H_{\alpha}} \sum_{s=1}^{M_{\alpha i}} X_{\alpha i s}$$

where $X_{\alpha i s}$ is the total of the characteristic in the s^{th} Master Sample segment in the i^{th} M.C.D. in the α^{th} stratum, $M_{\alpha i}$ is the total number of Master Sample segments in the i^{th} M.C.D. of the α^{th} stratum, H_{α} is the total number of M.C.D.'s in the α^{th} stratum, and G is the total number of strata. We shall consider three different estimates of X in this investigation.

(i) Unbiased Estimate The unbiased estimate can be expressed as

$$(9) \quad X' = \frac{1}{t} \sum_{\alpha=1}^G \sum_{i=1}^{H_{\alpha}} \sum_{s=1}^{m_{\alpha i}} X'_{\alpha i s}$$

where t is the over-all sampling rate and $m_{\alpha i}$ is the number of segments in the sample for the i^{th} M.C.D. of the α^{th} stratum. Thus in equation form

$$m_{\alpha i} = \frac{t P_{\alpha}}{P_{\alpha i}} \cdot M_{\alpha i},$$

where P_{α} is the total number of farms in the α^{th} stratum in 1940 and $P_{\alpha i}$ is the total number of farms in the i^{th} county of the α^{th} stratum in 1940.

The variance of the unbiased estimate, as derived in the appendix is,

$$(10) \sigma_{\bar{X}}^2 = \sum_{\alpha=1}^G \left[\frac{H_{\alpha} P_{\alpha i}}{\sum_{i=1}^H P_{\alpha}} \cdot \frac{n_{\alpha i}}{t} \cdot \frac{M_{\alpha i} - m_{\alpha i}}{M_{\alpha i} - 1} \sigma_{\alpha i}^2 \right] + \sum_{\alpha=1}^G \left[P_{\alpha}^2 \frac{H_{\alpha} P_{\alpha i}}{\sum_{i=1}^H P_{\alpha}} (\bar{X}_{\alpha i} - \bar{X}_{\alpha})^2 \right]$$

where

$$\bar{X}_{\alpha i} = \frac{\text{1945 characteristic in the } i^{\text{th}} \text{ M.C.D. of the } \alpha^{\text{th}} \text{ stratum}}{\text{1940 number of farms in the } i^{\text{th}} \text{ M.C.D. of the } \alpha^{\text{th}} \text{ stratum}}$$

$$\bar{X}_{\alpha} = \frac{\text{1945 characteristic in the } \alpha^{\text{th}} \text{ stratum}}{\text{1940 number of farms in the } \alpha^{\text{th}} \text{ stratum}}$$

$$\sigma_{\alpha i}^2 = \text{variance within the } i^{\text{th}} \text{ M.C.D. of the } \alpha^{\text{th}} \text{ stratum.}$$

Only the second part of (10) has been computed in this investigation. This part is called the "between primary sampling unit component" of the total variance.

(ii) Ratio Estimate To estimate \bar{X} we can also use the ratio estimate, which may be slightly biased. This estimate is

$$(11) \bar{X}'_r = Y \cdot \frac{\bar{X}'}{\bar{Y}'}$$

where \bar{X}' is an unbiased estimate of the population total for some specified characteristic and is given by

$$(12) \bar{X}' = \frac{1}{t} \sum_{\alpha=1}^G n_{\alpha i} \bar{X}_{\alpha i}$$

and Y and Y' are the population total and sample total of an auxiliary variate that is correlated with X' .

An approximation of the variance of the ratio estimate is given by the equation

$$(13) \quad \sigma_{X'_r}^2 \approx X^2 \left(\frac{\sigma_{X'}^2}{X^2} + \frac{\sigma_{Y'}^2}{Y^2} - \frac{2 \sigma_{X'Y'}}{XY} \right).$$

Thus the coefficient of variation of the ratio estimate is approximated by

$$(14) \quad (C.V._{X'_r})^2 \approx \frac{\sigma_{X'}^2}{X^2} + \frac{\sigma_{Y'}^2}{Y^2} - \frac{2 \sigma_{X'Y'}}{XY}$$

The bias of the ratio estimate is approximated by

$$(15) \quad \frac{X}{Y} (C.V._Y^2 - C.V._{X'Y'})^2.$$

The derivations of the bias, variance and $(C.V.)^2$ of the ratio estimate are presented in the appendix.

(iii) Regression Estimate The estimation equation for a population total, \underline{X} , for the regression estimate can be expressed as

$$(16) \quad X'_{Lr} = X' - \frac{\sigma_{X'Y'}}{\sigma_{Y'}^2} (Y' - Y)$$

where X' and Y' are unbiased estimates of the population total.

The variance of the regression estimate as derived in the appendix, is

$$(17) \quad \sigma_{X'_{Lr}}^2 = \sigma_{X'}^2 - \frac{(\sigma_{X'Y'})^2}{\sigma_{Y'}^2}$$

Table 4 presents the between primary sampling unit component of the (C.V.)² for three different items estimated for 1945. Two sets of comparisons have been made: (i) comparison of three different methods of estimation, (ii) comparison of five different types of stratification. Little difference can be observed between the unbiased and ratio estimates. As the unbiased estimate is easier to compute, it is recommended over the ratio estimate. In all cases the regression estimate showed slight gains over the unbiased estimate. It is not believed that this slight gain is worth the added costs of computing the regression coefficient. The most important conclusion that can be drawn from Table 4 is that the geographic stratification is better than the other four types considered in all cases for the unbiased estimate. In an actual survey, geographic stratification is not only easier to use from a computational point of view but also information on the stratification items is not necessary.

Table 5 shows the bias of the ratio estimate in per cent. The largest bias obtained was approximately 1/2 of 1 per cent; so for practical purposes this bias is negligible and can be neglected in the comparison of the ratio with the unbiased estimate.

In comparing the stratification by current Census information with that of past Census information, it was found that there is an improvement in using current information over past information. Information in 1940 and 1945 was available for only two of the stratification items. The results appear in Table 6. For "Days Worked Off Farms" the stratification by current Census information gave a higher (C.V.)² than for past Census information. This may be explained by the fact that this characteristic was more variable in 1945 than in 1940.

Table 4. Comparison of the Between Primary Sampling Unit Component of the (Coefficient of Variation)² for Approximately 200 Strata of M.C.D.'s Using Different Methods of Estimation.
1/

Estimate Used	Estimate of									
	Tobacco acreage in 1945	Cotton acreage in 1945	Days worked off farms in 1945	Ratio to :	Un-biased: of farms :	Un-biased: of farms :	Un-biased: of farms :	Un-biased: of farms :	Un-biased: of farms :	Un-biased: of farms :
Type of stratification:	:	:	:	:	:	:	:	:	:	:
Geographic	:	:	:	:	:	:	:	:	:	:
Days worked off farms in 1940	.645	.722	.595	1.150	1.329	1.132	4.920	5.620	4.910	5.040
Value of products sold in 1940	.931	.933	.881	3.147	3.139	3.092	5.616	5.135	5.632	6.048
Average nilos from all weather roads in 1945	1.016	.940	.917	1.528	1.323	1.271	6.083	5.674	5.632	6.048
Farms operated by tenants in 1945	1.016	1.009	.963	3.196	3.171	3.130	6.426	6.071	6.048	5.467
	1.016	1.030	.973	2.895	2.940	2.865	6.141	5.608	5.467	

1/ ALK (C.V.)² in this table have been multiplied by 10.³

Table 5. Bias of Ratio Estimate in Per Cent

1945 Estimate	Tobacco acreage in 1945	Cotton acreage in 1945	Days worked off farm in 1945
<u>Type of stratification</u>			
Geographic	.047	.064	.412
Days worked off farms in 1940	.023	.024	.449
Value of products sold in 1940	.016	.027	.308
Average miles from all weather roads in 1945	.020	.024	.023
Farms operated by tenants in 1945	.024	.030	.544

Table 6. Comparison of the Between Primary Sampling Unit Component of the (Coefficient of Variation)² for Stratification by Current Census Information with that for Past Census Information Using the Unbiased Estimate.

Type of stratification	1945 Estimate		
	Tobacco acres	Cotton acres	Days worked off farms
Days worked off farms in 1940	.000931	.003147	.005616
Days worked off farms in 1945	.000877	.003044	.005913
Value of products sold in 1940	.001016	.001329	.006083
Value of products sold in 1945	.001104	.001274	.005728

For the efficiency of the county and the M.C.D. as a primary sampling unit, several considerations must be made in comparing estimates based on our sampling plan with estimates obtained by Madow [1] . These considerations include gains due to stratification, cost, and size of primary sampling unit. We assume, as Madow does, that the between primary sampling unit variance is the dominant term in the total variance. It is also reasonable to assume that the between segment within county variance is at least as large as the between segment within M.C.D. variance. This becomes clear when the scatter of selected Master Sample segments within the county is compared with the scatter of selected segments within the M.C.D.

It can be seen from Table 7 that the between component is about twice as large for the M.C.D. with 20 strata as for the county with 20 strata. The difficulty in making this comparison is that the within variances are not known, and no consideration is made as to the relative costs of each method. It would seem that the county is more efficient than the M.C.D. using 20 strata. The reasoning follows:

(i) If the number of sub-samples is the same in each case, it is believed that the costs would be practically the same. The reasoning here is that the between primary sampling unit travel would be about the same since 20 are visited in each case. The between sub-unit travel within the primary sampling unit would slightly favor the M.C.D., but this would not be very great as the same number of sub-units must be visited.

(ii) The within variance for M.C.D.'s would be reduced by two factors:

- (a) elimination of the between M.C.D. variation within county
- (b) the finite population correction would reduce the variance more for the M.C.D., as the sampling rate would be higher.

However if we assume that the between component is dominant as was assumed

Table 7. Comparison of the (Coefficient of Variation)² of 20 Strata where the County is the Primary Sampling Unit and 20 and 197 Strata where the M.C.D. is the Primary Sampling Unit.^{1/}

P.S.U.	(Coefficient of Variation) ²									
	County	M.C.D.	M.C.D.	M.C.D.	M.C.D.	M.C.D.	M.C.D.	M.C.D.	M.C.D.	% Gain of 197 M.C.D.
Number of strata	2/	3/	4/	5/	6/	7/	8/	9/	10/	11/
Estimate used	Unbiased:	Ratio:	Unbiased:	Ratio:	Unbiased:	Ratio:	Unbiased:	Ratio:	Unbiased:	Ratio:
Estimated item	1	2	3	4	5	6	7	8		
Tobacco acreage	.0058	.0061	.0127	.0125	.000645	.000722	.49	.42		
Cotton acreage	.0182	.0184	.0324	.0263	.001150	.001329	.65	.49		
Days worked off farm:	.0231	.0205	.0596	.0574	.004920	.005620	.28	.23		

1/ Only the between primary sampling unit component is considered.
 2/ The figures in columns (1) and (2) are those obtained by Mrs. Meadow [1], where 20 strata were used and the county was the primary sampling unit.
 3/ Same strata as used in (1) and (2) but with the Minor Civil Division as the primary sampling unit.
 4/ Those figures are those obtained using 197 strata where the Minor Civil Division is the primary sampling unit.
 5/ Gain in per cent for 197 M.C.D. over 20 M.C.D. was calculated by:

$$\frac{1}{10} (C.V._{20})^2 - (C.V._{197})^2 \times 100$$

in [1] , it is not believed that the reductions mentioned in (a) and (b) would compensate for the differences between columns 1 and 3 and columns 2 and 4.

To compare the method where one county is selected from each of 20 strata with the method where one M.C.D. is selected from each of 197 strata is even more difficult without information on the within variances and relative costs. If an equal number of sub-units are selected, the sampling rate can be assumed to be the same. Thus the within variance for the M.C.D.'s would be reduced by only the factor (ii) (a) above. However it should be less than that for the within county. Thus it would seem reasonable for the 197 strata to have a smaller total variance than the 20 strata. As it is more costly to enumerate segments in 197 M.C.D.'s than for the same number of segments in 20 counties, it is not possible to make a definite statement as to the relative efficiencies of these two methods. However it seems reasonable that this comparison might be in favor of 197 strata.

In order to calculate the relative gains due to stratification where the size of the primary sampling unit is the same, we compare columns 3 and 4 with columns 5 and 6 in Table 7. The per cent gain of 197 strata over 20 strata where the M.C.D. was used as the primary sampling unit is indicated in columns 7 and 8. Considerable gains are indicated on all items for both the simple unbiased estimate and the ratio estimate. Greater gains were obtained for the simple unbiased estimate than for the ratio estimate. Increased stratification was more effective than the use of a more complicated estimation equation, neglecting costs.

At the present time research is being done on the within M.C.D. and county contribution to the over-all variances. When these calculations are complete, it will be possible to make more definite statements about

the efficiency of the M.C.D. and the county as a primary sampling unit. Little data are available on costs at present, but it is hoped to obtain such data in the near future.

APPENDIX

Pertinent notation for this thesis is presented in tabular form below.

	<u>State</u>	<u>Stratum</u>	<u>M.C.D. in stratum</u>	<u>Sampling unit in M.C.D.</u>
Subscript		α	αi	$\alpha i s$
Number in population	1	G	H_{α}	$M_{\alpha i}$
Number in sample	1	G	1	$n_{\alpha i}$
Total number of farms in 1940	P	P_{α}	$P_{\alpha i}$	$P_{\alpha i s}$
Total of agricultural characteristics in 1945	X	X_{α}	$X_{\alpha i}$	$X_{\alpha i s}$
Estimate of total	X'	X'_{α}	$X'_{\alpha i}$	$X'_{\alpha i s}$
<u>Total of characteristics in 1945</u>	\bar{X}	\bar{X}_{α}	$\bar{X}_{\alpha i}$	
Total number of farms in 1940				
Mean per sampling unit			$\bar{X}_{\alpha i} = \frac{X_{\alpha i}}{M_{\alpha i}}$	

Other Notation

t = sampling rate

$$X = \sum_{\alpha} X_{\alpha} = \sum_{\alpha} \sum_{i=1}^G X_{\alpha i} = \sum_{\alpha} \sum_{i=1}^G \sum_{s=1}^{n_{\alpha i}} X_{\alpha i s}$$

Proof that X' is unbiased estimate of X .^{1/}

$$E X' = E \left[\frac{1}{t} \sum_{\alpha=1}^G \sum_{i=1}^G \sum_{s=1}^{n_{\alpha i}} X'_{\alpha i s} \right] = \frac{1}{t} \sum_{\alpha=1}^G E_{\alpha} \left[\sum_{i=1}^G \sum_{s=1}^{n_{\alpha i}} E_{\alpha i} X'_{\alpha i s} \right]$$

^{1/} This proof is presented in reference [5] .

where E_{α} is the expected value for the α^{th} stratum and $E_{\alpha i}$ is the expected value of $X'_{\alpha i s}$ within the county selected from the α^{th} stratum.

$$\text{Now, } E_{\alpha i} X'_{\alpha i s} = \frac{1}{M_{\alpha i}} \sum_{s=1}^{M_{\alpha i}} X_{\alpha i s} = \bar{X}_{\alpha i}$$

Then,

$$E_{\alpha} \left[\frac{1}{\sum_{i=1}^{n_{\alpha i}} \sum_{s=1}^{M_{\alpha i}} E_{\alpha i} X'_{\alpha i s}} \right] = E_{\alpha} \left[\frac{1}{\sum_{i=1}^{n_{\alpha i}} \sum_{s=1}^{M_{\alpha i}} \bar{X}_{\alpha i}} \right] = E_{\alpha} \sum_{i=1}^{n_{\alpha i}} n_{\alpha i} \bar{X}_{\alpha i}$$

To get $\bar{X}_{\alpha i}$, or the 1945 mean per 1940 number of farms, remember that

$$n_{\alpha i} = \frac{t P_{\alpha}}{P_{\alpha i}} M_{\alpha i}$$

Thus,

$$n_{\alpha i} \bar{X}_{\alpha i} = \frac{t P_{\alpha}}{P_{\alpha i}} M_{\alpha i} \cdot \bar{X}_{\alpha i} = t P_{\alpha} \tilde{X}_{\alpha i}$$

Therefore,

$$E X' = \sum_{\alpha=1}^G E_{\alpha} \left[\frac{1}{\sum_{i=1}^{n_{\alpha i}} P_{\alpha i} \tilde{X}_{\alpha i}} \right] = \sum_{\alpha=1}^G P_{\alpha} \sum_{i=1}^{n_{\alpha i}} E_{\alpha} \tilde{X}_{\alpha i}$$

Now,

$$E_{\alpha} \tilde{X}_{\alpha i} = \sum_{i=1}^{n_{\alpha i}} \frac{P_{\alpha i}}{P_{\alpha}} \tilde{X}_{\alpha i} = \tilde{X}_{\alpha}$$

where $\frac{P_{\alpha i}}{P_{\alpha}}$ = the probability of selecting the i^{th} county from the α^{th} stratum, using probability proportional to the 1940 number of farms.

Thus,

$$E X' = \sum_{\alpha=1}^G P_{\alpha} \tilde{X}_{\alpha} = \sum_{\alpha=1}^G \sum_{i=1}^{n_{\alpha i}} \sum_{s=1}^{M_{\alpha i}} X_{\alpha i s} = X$$

Derivation of Variance of Unbiased Estimate.^{1/} We may write $X^i = \sum_{\alpha=1}^G X_{\alpha}^i$,

where $X_{\alpha}^i = \frac{1}{t} \sum_{i=1}^1 \sum_{s=1}^n c_{\alpha is} X_{\alpha is}^i$

Then

$$E X^i = \sum_{\alpha=1}^G E X_{\alpha}^i,$$

and if the sample is selected independently from each stratum, we have

$$\sigma_{X^i}^2 = \sum_{\alpha=1}^G \sigma_{X_{\alpha}^i}^2$$

Thus, we need to evaluate $\sigma_{X_{\alpha}^i}^2$. By the definition of variance, we have

$$(1) \dots \sigma_{X_{\alpha}^i}^2 = E \left[X_{\alpha}^i - E X_{\alpha}^i \right]^2 = E_{\alpha} E_{ci} \left[X_{\alpha}^i - E X_{\alpha}^i \right]^2$$

We may write

$$(2) \dots E_{ci} \left[X_{\alpha}^i - E X_{\alpha}^i \right]^2 = E_{ci} \left[X_{\alpha}^i - E_{ci} X_{\alpha}^i \right]^2 + \left[E_{ci} X_{\alpha}^i - E X_{\alpha}^i \right]^2,$$

since

$$E_{ci} \left[X_{\alpha}^i - E_{ci} X_{\alpha}^i \right] \left[E_{ci} X_{\alpha}^i - E X_{\alpha}^i \right] = \left[E_{ci} X_{\alpha}^i - E X_{\alpha}^i \right] \cdot E_{ci} \left[X_{\alpha}^i - E_{ci} X_{\alpha}^i \right] = \text{constant } X \left[E_{ci} X_{\alpha}^i - E_{ci} X_{\alpha}^i \right] = 0$$

From (1) and (2), we then have

$$E \left[X_{\alpha}^i - E X_{\alpha}^i \right]^2 = E_{\alpha} \left[E_{ci} \left(X_{\alpha}^i - E_{ci} X_{\alpha}^i \right)^2 \right] + E_{\alpha} \left[E_{ci} \left(E_{ci} X_{\alpha}^i - E X_{\alpha}^i \right)^2 \right] = E_{\alpha} (I) + E_{\alpha} (II)$$

^{1/} This proof is presented in reference [5].

To evaluate $E_{\alpha} (I)$, we calculate

$$E_{\alpha i} \left[X'_{\alpha} - E_{\alpha i} X'_{\alpha} \right]^2 = \frac{1}{t^2} m_{\alpha i} \frac{M_{\alpha i} - m_{\alpha i}}{M_{\alpha i} - 1} \cdot \sigma_{\alpha i}^2,$$

where

$$\sigma_{\alpha i}^2 = \frac{1}{M_{\alpha i}} \left[\sum_{s=1}^{M_{\alpha i}} (X'_{\alpha i s} - \bar{X}_{\alpha i})^2 \right]$$

Thus,

$$E_{\alpha} (I) = E_{\alpha} \left[\frac{1}{t^2} m_{\alpha i} \frac{M_{\alpha i} - m_{\alpha i}}{M_{\alpha i} - 1} \sigma_{\alpha i}^2 \right] = \frac{H}{\sum_{\alpha}^{\alpha} P_{\alpha}} \frac{P_{\alpha i}}{P_{\alpha}} \cdot \frac{m_{\alpha i}}{t^2} \cdot \frac{M_{\alpha i} - m_{\alpha i}}{M_{\alpha i} - 1} \sigma_{\alpha i}^2$$

To evaluate

$$E_{\alpha} (II) = E_{\alpha} \left[E_{\alpha i} X'_{\alpha} - E X'_{\alpha} \right]^2,$$

we note that

$$E X'_{\alpha} = X_{\alpha} = P_{\alpha} \tilde{X}_{\alpha}$$

Also

$$E X'_{\alpha} = E_{\alpha} (E_{\alpha i} X'_{\alpha})$$

and

$$E_{\alpha i} X'_{\alpha} = P_{\alpha} \tilde{X}_{\alpha i}$$

Thus,

$$\begin{aligned} E_{\alpha} (II) &= E_{\alpha} \left[P_{\alpha} X_{\alpha i} - P_{\alpha} \tilde{X}_{\alpha} \right]^2 = P_{\alpha}^2 E_{\alpha} \left[\tilde{X}_{\alpha i} - \tilde{X}_{\alpha} \right]^2 \\ &= P_{\alpha}^2 \frac{H}{\sum_{\alpha}^{\alpha} P_{\alpha}} \frac{P_{\alpha i}}{P_{\alpha}} (\tilde{X}_{\alpha i} - \tilde{X}_{\alpha})^2 \end{aligned}$$

Thus, we have, finally

$$\begin{aligned} \sigma_{X'}^2 &= \sum_{\alpha=1}^G \sigma_{X'_{\alpha}}^2 = \sum_{\alpha=1}^G \left[\frac{H}{\sum_{\alpha}^{\alpha} P_{\alpha}} \frac{P_{\alpha i}}{P_{\alpha}} \cdot \frac{m_{\alpha i}}{t^2} \cdot \frac{M_{\alpha i} - m_{\alpha i}}{M_{\alpha i} - 1} \cdot \sigma_{\alpha i}^2 \right] \\ &+ \sum_{\alpha=1}^G \left[P_{\alpha}^2 \frac{H}{\sum_{\alpha}^{\alpha} P_{\alpha}} \frac{P_{\alpha i}}{P_{\alpha}} (X_{\alpha i} - X_{\alpha})^2 \right] \end{aligned}$$

where

$$\frac{1}{t^2} = \left[\frac{P}{P_{ci}} \cdot \frac{M_{ci}}{m_{ci}} \right]^2$$

Derivation of the Variance of the Ratio Estimate.^{1/}

The ratio estimate of X is

$$X'_r = Y \cdot \frac{X'}{Y'}$$

where X' is an unbiased estimate of the population

total of a characteristic and Y and Y' are the population total and sample total of an auxiliary variate that is correlated with X.

The variance of the ratio estimate can be approximated as follows.

$$X'_r = Y \cdot \left(\frac{X + \Delta X}{Y + \Delta Y} \right) \text{ where } X' = X + \Delta X$$

$$Y' = Y + \Delta Y$$

This can be written as

$$X'_r = X \left(1 + \frac{\Delta X}{X} \right) \left(1 + \frac{\Delta Y}{Y} \right)^{-1} \text{ and expanding the last term}$$

$$X'_r = X \left(1 + \frac{\Delta X}{X} \right) \left(1 - \frac{\Delta Y}{Y} + \dots + \frac{(\Delta Y)^n}{Y^n} \right) \text{ and eliminating terms}$$

of order $\frac{1}{Y^2}$ etc. the expression can be written as

$$X'_r \cong X \left[\frac{\Delta X}{X} + 1 - \frac{\Delta Y}{Y} \right]$$

Since

$$E X'_r \cong X,$$

$$X'_r - E X'_r \cong X \left[\frac{\Delta X}{X} - \frac{\Delta Y}{Y} \right], \text{ neglecting the term in } (\Delta X)(\Delta Y).$$

^{1/} This proof is presented in reference [4].

Also

$$E (X'_r - E X'_r)^2 \approx X^2 E \left(\frac{\Delta X}{X} - \frac{\Delta Y}{Y} \right)^2 \approx X^2 \frac{\sigma_{X'}^2}{X^2} + \frac{\sigma_{Y'}^2}{Y^2} - \frac{2 \sigma_{X'Y'}}{XY}$$

$$(C.V. X'_r)^2 \approx \frac{\sigma_{X'}^2}{X^2} + \frac{\sigma_{Y'}^2}{Y^2} - \frac{2 \sigma_{X'Y'}}{XY}$$

Derivation of the Bias of the Ratio Estimate:^{1/}

The estimation equation as shown previously is

$$X'_r = Y \cdot \frac{X'}{Y'}$$

Since $E \frac{X'}{Y'} \neq \frac{E X'}{E Y'}$ the approximate bias can be found by expanding $\frac{X'}{Y'}$ in an infinite series about $E X'$ and $E Y'$.

Thus

$$E \frac{X'}{Y'} - \frac{X}{Y} = E \left(\frac{X'}{Y'} \right) - \frac{X}{Y}$$

This can be expressed as

$$E \left(\frac{X'}{Y' - E Y' + E Y'} \right) - \frac{X}{Y}$$

$$= E \left(\frac{\frac{X'}{E Y'}}{1 + \frac{Y' - E Y'}{E Y'}} \right) - \frac{X}{Y}$$

$$= E \left(\frac{X'}{E Y'} \right) \left(1 - \frac{Y' - E Y'}{E Y'} + \frac{(Y' - E Y')^2}{(E Y')^2} - \dots \right) - \frac{X}{Y}$$

Therefore

$$E \left(\frac{X'}{Y'} - \frac{X}{Y} \right) = \left(\frac{E X'}{E Y'} - \frac{X}{Y} \right) - \left(\frac{E X' (Y' - E Y')}{(E Y')^2} \right) + \left(\frac{E X' (Y' - E Y')^2}{(E Y')^3} \right) - \dots$$

^{1/} This proof is presented in reference [5] .

Since $E X' = X$ and $E Y' = Y$, the first term vanishes and

$$\left(E \frac{X'}{Y'} - \frac{X}{Y} \right) = - \left(\frac{E X' (Y' - E Y')}{Y^2} \right) + \left(\frac{E X' (Y' - E Y')^2}{Y^3} \right) - \dots$$

The first term is the covariance and can be expressed as $-\frac{\sigma_{X'Y'}}{Y^2}$.

Adding and subtracting $\frac{E X' E(Y' - E Y')^2}{Y^3}$, the second term may be written as

$$E \left(\frac{(X' - E X')(Y' - E Y')^2}{Y^3} \right) + \frac{E X' E(Y' - E Y')^2}{Y^3}$$

The first term is of order $\frac{1}{Y^2}$ and can be discarded.

The second can be written as

$$\frac{X \sigma_{Y'}^2}{Y^3}$$

An approximation to the bias can then be written as

$$\left(E \frac{X'}{Y'} - \frac{X}{Y} \right) = - \frac{\sigma_{X'Y'}}{Y^2} + \frac{X \sigma_{Y'}^2}{Y^3}$$

$$= - \frac{X}{Y} \text{C.V.}_{Y'}^2 + \text{C.V.}_{X'Y'}$$

Derivation of the Variance for the Regression Estimate: ^{1/}

The estimation equation for a population total, X, for the regression estimate can be expressed as

$$X'_{Lr} = X' - \frac{\sigma_{X'Y'}}{\sigma_{Y'}^2} (Y' - Y) \text{ where } X' \text{ and } Y' \text{ are unbiased estimates of}$$

^{1/} This proof is presented in reference [5].

the population total and thus $E X' = E X'_{Lr} = X$ and $E Y' = Y$

so $\sigma_{X'_{Lr}}^2 = E(X'_{Lr} - X)^2$ and substituting the following results

$$E \left[(X' - X) - \frac{\sigma_{X'Y'}}{\sigma_{Y'}^2} (Y' - Y) \right]^2 \quad \text{and taking the expected value the}$$

following is found

$$\sigma_{X'}^2 - \frac{2(\sigma_{X'Y'})^2}{\sigma_{Y'}^2} + \sigma_{Y'}^2 \frac{(\sigma_{X'Y'})^2}{(\sigma_{Y'}^2)^2} \quad \text{which is equal to}$$

$$\sigma_{X'_{Lr}}^2 = \sigma_{X'}^2 - \frac{(\sigma_{X'Y'})^2}{\sigma_{Y'}^2}$$

REFERENCES

- (1) Madow, Lillian H. "On Sampling for State Agricultural Estimates in North Carolina" Journal of the American Statistical Association. Volume 45. March 1950.
- (2) King, A. J. and Jessen, R. J. "The Master Sample of Agriculture." Journal of the American Statistical Association. Volume 40. March 1945.
- (3) Hansen, Morris H. and Hurwitz, William N. "On the Theory of Sampling from Finite Populations." The Annals of Mathematical Statistics. Volume XIV. No. 4. Page 333. December 1943.
- (4) Cochran, W. G. "Sample Survey Techniques." Mimeographed by the Department of Statistics at North Carolina State College. June 1948.
- (5) Madow, Lillian H. "On Sampling for State Agricultural Estimates in North Carolina." Progress Report, North Carolina State College. February 1949.