

Miss Cox

A HANDBOOK OF AREA SAMPLING

by

John Monroe and A. L. Finkner

November, 1957

Institute of Statistics
Mimeo Series No. 185

TABLE OF CONTENTS

| | Page |
|--|------|
| FOREWORD | |
| I INTRODUCTION | 1 |
| II PREPARATION OF THE SAMPLING MATERIALS | 3 |
| A. Use of The Census Data | 3 |
| B. The Open Country Count Map | 4 |
| C. Materials for Urban and Rural Places | 7 |
| III USE OF THE MATERIALS | 9 |
| A. Size of The Sampling Unit | 9 |
| B. Allocation of The Sampling Units | 10 |
| C. The Sample Draw | 13 |
| 1. The Urban Sample | 14 |
| 2. The Rural Plan Sample | 21 |
| 3. The Open Country Sample | 25 |
| IV SOME STANDARD SAMPLE DESIGNS | 28 |
| A. Simple Random Sampling | 28 |
| B. Stratified Random Sampling | 29 |
| 1. General Case - Unequal Sized Strata | 30 |
| 2. Special Case - Equal Sized Strata | 32 |
| 3. Optimum versus Proportional Allocation | 34 |
| C. Two-Stage Sampling | 35 |
| 1. Selection of The First-Stage Units with Unequal Probabilities | 36 |
| a. One Sample FSU per Stratum | 36 |
| b. Two or More Sample FSU per Stratum | 37 |
| 2. Selection of The First Stage Units with Equal Probability | 39 |

REFERENCES

APPENDIX

| | |
|-------------------------|---|
| A. Census Definitions | 1 |
| B. Housing Data Sources | 4 |
| C. Aerial Photo Service | 5 |

FOREWORD

Since the development of the Master Sample of Agriculture sampling materials in 1943-1945, the use of "area" sampling in surveys has become widespread in a variety of interests - non-agricultural as well as agricultural - and types - analytical as well as descriptive.

Availability of area sampling materials in federal, state, and private agencies has contributed much to the awareness of the use of area sampling. Demand for area sample design, consultation, and preparation appears to have reached the point where the researcher's appreciation of the importance of a probability sample has outrun his cognizance of the problems encountered in defining the universe of the sample, the organization, or frame, of the universe, the reliability of the information required for the sample allocation and draw, and the availability of the materials for the designation of the sample for field use. This failing is due in part to the recognition of the "total picture" economy where we find "agricultural" researchers working in urban areas and "urban" researchers trying to determine such things as the rural contributions to the industrial labor force. In 1955, for example, a regional committee of agricultural economists conducted a study of milk consumption wholly within 12 metropolitan cities in the Southeast.

Unfortunately, there still exists among researchers the idea that drawing a sample is a mechanical, "drawer-pulling-out" process by those having access to mountainous files of maps, punched-cards, and copies of Census publications. There is also a tendency to call upon experts to do the sample work without exploring the possibilities

of constructing the sample frame "at home," where clerical and supervisory man-power might be available and the more expensive materials at hand.

In the process of developing sampling materials for North Carolina and, at the same time, fulfilling requests for samples in other states, the writers found it necessary to explain repeatedly the process of constructing an area sample frame. This explanation included the types and sources of information and maps for the sample allocation and draw, the office and field work required for the definition and delineation of the sampling unit and observation unit, and the estimated costs of the various steps in the sampling procedure. In many instances it appeared that in a value vs. cost concept the researcher should restrict or re-define his universe of inquiry within the existing available materials, or actually perform at least part of the construction of the sample frame himself. Frequently, assembling the materials to meet the sample specifications would cost more than the total survey budget!

The purpose of this pamphlet is twofold: to bring together the various definitions and procedures involved in the construction of an area sample frame and to indicate the use of the frame in drawing samples. The writers believe that a detailed discussion of procedures will be informative to those who might be in a position to make a collection of sampling materials for continuous use and that appreciation of this operation will be beneficial to the researcher who makes use of sample surveys for data collection.

The authors do not in any way consider the concepts and practices set forth in this publication as original with them. We present this pamphlet as a collection of ideas and suggestions which have arisen during experience with surveys using the Master Sample of Agriculture

sampling materials and from association with sampling people throughout the country who have long felt the need to record the trials, improvisations, and conversations relevant to area sampling.

Among others, we gained enlightenment from R. L. Anderson, W. G. Cochran, W. E. Deming, J. Fleischer, M. H. Hansen, H. O. Hartley, W. A. Hendricks, D. G. Horvitz, E. E. Houseman, W. N. Hurwitz, E. M. Jacobs, R. J. Jessen, A. J. King, R. K. McMillan, T. J. Reed, A. Ross, N. V. Strand, and D. J. Thompson.

I INTRODUCTION

The first requirement for a probability sample of any nature is the establishment of a sampling frame. A sampling frame is a collection of sampling units which may be unambiguously defined and identified. A list of persons, families, or houses might be a sufficient frame for certain types of samples; for other samples, a list may not exist or cannot be inexpensively obtained. In the latter case, the Master Sample of Agriculture materials have provided a solution. These materials constitute a geographic frame of area units, "count" units, for the United States, whereby any element which has an association with a unit of area can be identified after locating a particular count unit. The count units vary in size and shape, as well as the "counts" - the number of farms and number of dwellings indicated on highway maps and aerial photographs. The one characteristic common to all is definable area. Within the count unit, sampling units can be defined and identified. This, in brief, is an area sampling frame. The frame thus constructed is entirely adequate for area probability sampling.

In the Master Sample materials, the actual allocation of a number of sampling units with certain characteristics in terms of number of farms and dwellings was made for each count unit - specifically for the Master Sample of Agriculture. Although at the time it was visualized that this allocation of sampling units to the count units would serve for many subsequent samples, re-assignments of sampling units consistent with the sample objectives have since been practiced. For example, a count unit which might have been assigned 2 sampling units for a farm study of farms over 50 acres in size might be assigned 4 sampling units for a study of males, 30 years of age and older.

Thus the count unit is a convenient device for recording

characteristics used in the assignment of sampling units. For a given sample, criteria of sampling units are applied to the count unit and the "size" of the count unit in terms of the number of sampling units is established. It is important to note that the ultimate sampling units need not be definable on a map, but might be defined at a subsequent stage in the survey procedure.

Experience with the Master Sample materials has provided incentive of North Carolina to develop new materials for the state/and other states when feasible. In addition to the obvious advantage of newer maps and culture, there remain two fundamental reasons for this development:

- 1) To provide a sampling frame which is consistent with the 1950 Census of Housing, and
- 2) To improve, by virtue of experience with the Master Sample materials, the efficiency of area probability samples.

In general, the present concept of the sampling materials follows that of the Master Sample of Agriculture. The main exceptions are:

- 1) ~~Abandonment~~ of the Minor Civil Division (MCD) as a geographic delimitation of count units. Only established roads, streams, railroads, and city, town, and county limits are now used as count unit boundaries. Since MCD housing data are not published and MCD boundaries are often difficult to locate in the field, there appears to be no advantage to preserving the practice at this time.
- 2) a) The URBAN ZONE now consists of urban places plus the urbanized areas defined in the 1950 Census.
b) The new definition of RURAL PLACE as all incorporated places less than 2500 in population and unincorporated places of 1000 to 2500 in population as defined by the Census.
c) The OPEN COUNTRY ZONE is the residual area not defined as Urban or Rural Place.
- 3) The unit of compilation is the occupied dwelling unit (ODU), or household. This unit is consistent with the Census data for all geographic areas through the county, urban places, and rural places with population 1000 to 2500. For incorporated towns less than 1000 in population, the number of ODUs is estimated by a population ratio.

II PREPARATION OF THE SAMPLING MATERIALS

The basic component of the sampling materials is the county 1" = 1 mile highway map. Map information and detail vary from state to state, yet these maps ordinarily show all cities, towns, the county road system, farm and non-farm dwelling units (vacant and occupied), stores, churches, schools, cemeteries, power lines, state parks, etc. The date of map reproduction and culture count are usually shown for each map.

A. Use of the Census Data

The 1950 Census of Population lists the names and populations of the cities and towns in each county. Classification of the places into Rural Place and Urban is thus determined. If the county contains an unincorporated place, an Enumeration District (ED) map is necessary to establish the Census delineated boundaries on the county map.

A city of 50,000 or more in population may mean the presence of urbanized areas around that city. Map and verbal descriptions of urbanized areas are found in both the Population and Housing volumes. The Census volume maps also show the 1950 city limits of the metropolitan city.

Although the county highway maps do show city and town limits, these limits are not necessarily the 1950 limits, and, in order to comply, the Census delineation must be used rather than the county delineation. To be completely accurate, ED maps for all cities and towns in the county should be used. The alternative to this, however, is less expensive: obtaining ED maps only for places as they are selected in the samples and using the county map boundaries for other incorporated places as indicated on the map. Compliance with the 1950 Census limits can then be dealt with in the field.

After the delineation of urban places, urbanized areas, and rural incorporated and unincorporated places, the residual area of the county is the Open Country. The 1950 number of ODUs in the Open Country is obtained by subtraction of the place totals from the county total. Buncombe County, North Carolina is used to illustrate the above procedure.

| 1950 CENSUS | | |
|--------------------------------------|---------------|----------------------------|
| BUNCOMBE COUNTY NORTH CAROLINA | POPULATION | OCCUPIED DWELLING UNITS |
| <u>URBAN</u> | | |
| Asheville | 53,000 | 15,029 |
| Urban Fringe | 5,437 | 1,455 |
| <u>RURAL PLACE</u> | | |
| Biltmore Forest | 657 | 178* |
| Black Mountain | 1,174 | 407 |
| Jupiter | 136 | 37* |
| Enka (uninc.) | 1,792 | 487 |
| Weaverville | 1,111 | 315 |
| Swannanoa-Grovemont (uninc.) | 1,913 | 523 |
| <u>TOTAL PLACES</u> | <u>65,220</u> | <u>18,431</u> |
| <u>OPEN COUNTRY (by Subtraction)</u> | <u>59,183</u> | <u>14,418</u> |

*For North Carolina, the average population/occupied dwelling unit in places of 1000 to 2,500 (Census information on occupied dwelling units is not published for towns under 1000 population) is 3.69. This average was used throughout the state to estimate the number of ODUs for towns under 1000 population.

B. The Open Country Count Map

In preparing the Buncombe county highway map, the first step is to delineate Asheville and its urban fringe. Delineation of the Census unincorporated places, Enka and Swannanoa-Grovemont, from the ED maps follows: and, finally, the marking of the incorporated places (in this county, they are all Rural Places) as indicated on the map

itself. There are map "name" places and insets (enlargements), not listed as Census places, which are considered part of the Open Country and included in the Open Country count units.

The organization of the Open Country will vary from state to state and between counties within state. In North Carolina, the main highways generally transect a county such that a spoked-wheel appearance is made with a town (usually the county seat) at the hub. Since main highways are relatively stable and easily identified, they serve as better boundaries for "divisions" than MCD lines. The division serves two purposes: convenience in recording the map counts and locating count units, and creation of areas which could be used as fairly homogeneous primary sampling units within counties.

The divisions are then marked in a distinctive color and numbered in a contiguous, counter-clockwise manner. Within each division, the main roads, rivers, and railroads provide stable and identifiable boundaries for the count units. The occupied dwelling units indicated on the map are then counted within each count unit. Since the count unit is merely a collection of sampling units, the number of which varies with the sample allocation, the size of each count unit, in terms of indicated number of dwellings (INOD), does not affect the ultimate selection of a sampling unit. However, the main road system and topography, together with the area extent, do present some limitations on the INOD size of the count unit. In sparsely settled areas, the few roads are the only available boundaries and make a count unit large in area but small in number of INOD. For operational purposes, the "target" size of the count unit is 15 INOD, with many exceptions to the rule. In some suburban areas, the size of the count unit may be 100 or more INOD simply because of the lack of adequate map boundaries. The practical minimum number of INOD is 3. If a count unit has less than 3 INOD it would probably be necessary to combine it with an adjacent count unit in order to

have a probability for **selection** as part of a sampling unit; this would defeat the purpose of the count unit.

An order of preference for count unit boundaries might be generally stated. The first principle, when choices exist, is to select the most stable boundary; e.g., a paved or surfaced road rather than a dirt road or a trail, a dirt road rather than a small stream which has no bridges to indicate road crossings. A large river, on the other hand, should be used as a count unit boundary, when possible, so that canvassing a possible sampling unit would not involve crossing and re-crossing a bridge which may be miles removed from the sample area. In mountainous and swampy areas, the road system is often inadequate for count unit boundaries and imaginary lines between identifiable land marks are necessary to form workable count units. Problems of travel for field personnel should be considered in the count unit delineations.

After the count units have been marked in color, they should be numbered contiguously such that the last numbered count unit of a division is adjacent to the first numbered count unit in the next division. With this procedure, the final list of count units is geographically as well as numerically ordered, and any combination of ordered count units can be made for purposes of stratification, primary sampling units, etc.

In some counties, the insets of congested areas help to improve the accuracy of the count of INOD. Occasionally, built-up places, especially suburban areas near cities do not have the dwellings indicated either on the maps or in the insets. For such areas as these, recent aerial photography can be used to obtain dwelling counts.

Experience with congested areas (which are roughly the equivalent of the Master Sample unincorporated stratum) may lead to a separate

identification of count units containing such areas. Congested areas usually mean a low number of farms or, at best, a number of small farms. Separate sampling of such a "stratum" would result in a more efficient sample for certain types of agricultural surveys.

The next step is the listing of the count units in numerical sequence by division, and count unit within division. Statistical (columnar) tabulating paper is useful for this purpose and provides a convenient record^{for} transfer of data to punched cards when desirable. The INOD is entered to the right of the count unit and division number, allowing space for accumulating totals on the listing sheet. Cumulative totals both within each division and over the entire county are helpful in the selection of samples. In the illustration of a selection technique below, the listing method is displayed.

C. Materials for Urban and Rural Places

A process similar to the Open Country method can be accomplished for the places in the county. However, as indicated earlier, savings are realized if the materials are ordered only as required for a specific sample.

Census Block Statistics are adequate for sampling cities with more than 50,000 population; the remaining places have to be dealt with individually. The ED maps supplied by the Census are sufficiently detailed to locate an ED within a place, which narrows the problem of obtaining block dwelling unit counts to sections of the place. The size of an ED varies considerably - in North Carolina the average is about 165 ODUs. Once the ED is selected, block counts may be made from aerial photos, Sanborn maps, town records, or by "cruising" the area for estimated counts. As a rule, Sanborn maps aren't readily available and town records are incomplete. The photo count and cruise methods are usually satisfactory for places of this

size; however, in apartment-house areas the photo-count method is subject to more variation than the cruise procedure. If the cruising scheme can be incorporated at the field stage, the cost is not much greater than the photo count. If time permits, the photo method is undoubtedly the cheaper, and the photos serve as maps for places where maps are unavailable or inadequate.

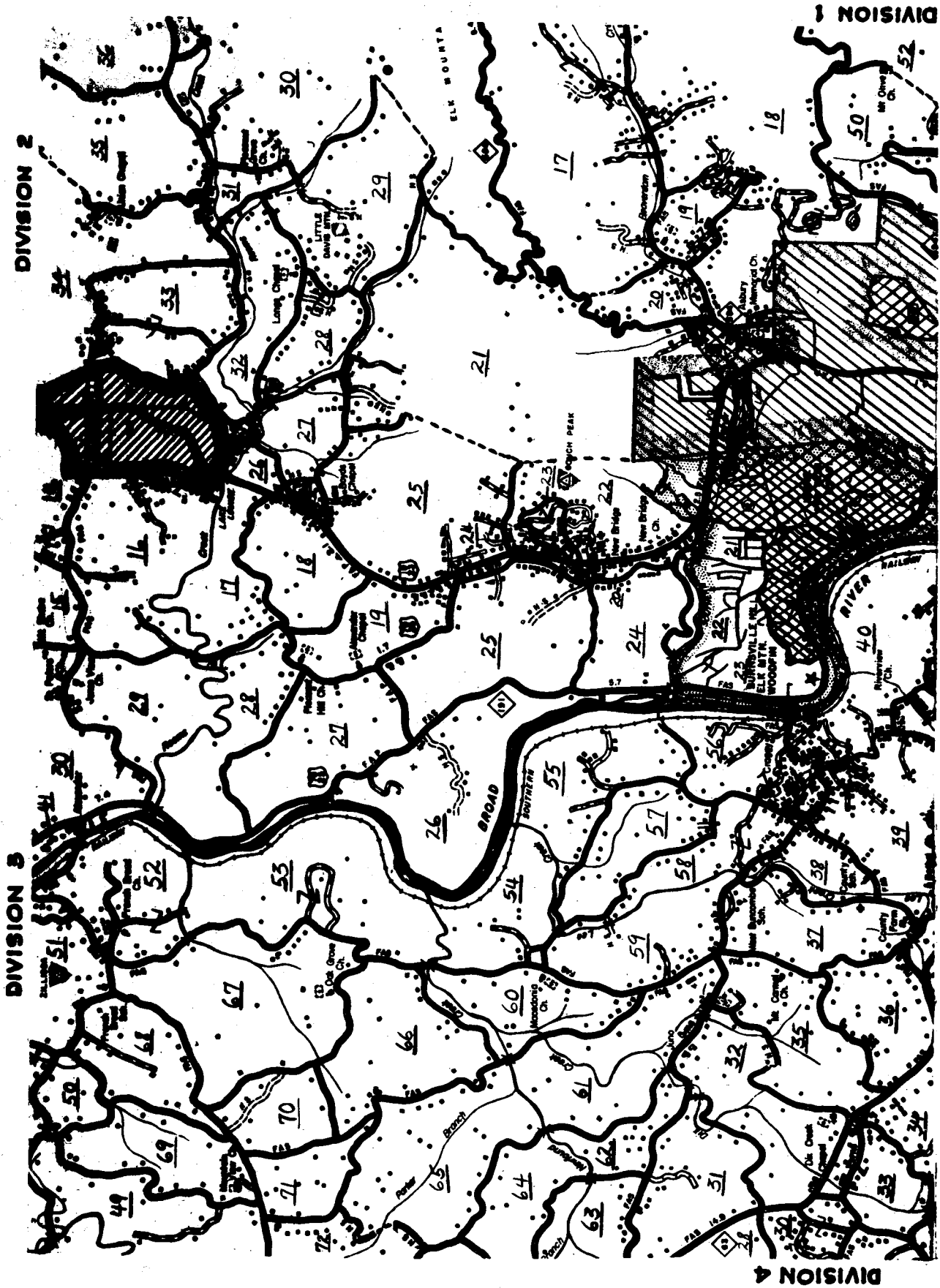


FIGURE 1: PORTION OF BUNCOMBE COUNTY COUNT MAP



III USE OF THE MATERIALS

In the development and application of an area frame, the sampling ~~materials discussed above are used essentially for two purposes:~~

- 1) To permit the determination of the number of sampling units to be assigned to a given area such that a probability sample may be selected. This implies the definition of the sampling unit in terms of the elements to be identified or observed (observation unit); e.g., households, farms, crops, individuals. (An area sampling unit per se is usually of no interest to the researcher; the unit of observation is the object of concern.)
- 2) To permit the identification in the field of a particular sampling unit which has been selected. (Each observation unit, in turn, must be identifiable with one, and only one, sampling unit. The area sampling unit is merely a convenient form of insuring that each observation unit has a known probability of selection).

The latest available information, usually published Census data, is used for the first purpose.¹ Block Statistics, Census Enumeration District maps and listings, aerial photographs, and delineated highway maps comprise materials for the second purpose, although they may also be used to some extent for the first.

Specific uses of the sampling materials for both purposes are discussed and illustrated step by step in the sampling procedure. The methods of identifying the selected sampling units is independent of the sample design and will be described only for a stratified random sample of households in Buncombe County, North Carolina.

A. Size of The Sampling Unit

If the sampling materials described are used, the size of the sampling unit can be made almost any size desired. The question the survey statistician must answer is which of a number of alternative sizes should be chosen. The answer to this question depends

¹ See Appendix I for Census definitions and sources of data.

upon a number of factors: the purpose of the survey, characteristics of the observation unit, the variability among sampling units of a given size, the sample design, the field work plan, and cost. The next step is to select a size of sampling unit in terms of an average (expected) number of observation units per area sampling unit.

In general, for a given size (bulk) of sample, a small unit is more precise than a large unit, yet it is also more expensive. If the optimum unit is defined as that which gives the greatest precision at a given cost, and if cost and variance estimates are available, the size of the optimum unit can be estimated. See Cochran (1953). However, in many situations the information necessary to make such a decision is not available and the selection has to be based upon experience. In general population surveys, the average size of the sampling unit usually varies between 3 and 6 households. The size of the sampling unit will vary in other types of surveys depending upon the purpose and available resources.

For this example, consideration of the factors affecting the size of the sampling unit led to the selection of an expected size of approximately four households per sampling unit. On the basis of 32,849 households in the universe, a sampling rate of about 1 in 100 should result in obtaining approximately 328 interviews - 82 sampling units, of expected size 4, in the sample.

B. Allocation of The Sampling Units

The number of sampling units assigned to a universe depends, of course, on the expected size of the sampling unit and the sample design.

In this example, the universe, Buncombe County, is stratified by the three zones; Urban, Rural Place, and Open Country. From Census information, the 1950 number of ODUs in Buncombe County

per stratum are:

| | |
|----------------------------------|---------------|
| Urban | 16,484 |
| Rural Place | 1,947 |
| Open Country (by subtraction) | <u>14,418</u> |
| Total | 32,849 |

Dividing the number of ODUs in each stratum by 4, the expected size of the sampling units, establishes the number of sampling units in the universe by stratum. In forcing the numbers of sampling units in the sample to be integral numbers in each stratum, the expected size of the sampling unit in each stratum now differs slightly from 4:

| <u>Stratum</u> | <u>Original Allocation</u> | | | <u>Adjusted Allocation</u> | |
|----------------|----------------------------|----------------------|----------------------------|----------------------------|----------------------------|
| | <u>Number of ODUs</u> | <u>Number of SUs</u> | <u>Expected Size of SU</u> | <u>Number of SUs</u> | <u>Expected Size of SU</u> |
| Urban | 16,484 | 4,121 | 4.0000 | 4,100 | 4.0205 |
| Rural Place | 1,947 | 487 | 3.9979 | 500 | 3.8940 |
| Open Country | <u>14,418</u> | <u>3,604</u> | 4.0006 | <u>3,600</u> | 4.0050 |
| Total | 32,849 | 8,212 | 4.0001 | 8,200 | 4.0060 |

The adjusted allocation is made to conform with the sampling rate of 1/100, i.e., the application of a 1/100 sampling rate will allow exactly an integral number of sampling units to be drawn in each stratum. Thus, 41 sampling units will be selected from the Urban stratum, 5 from the Rural Place stratum and 36 from the Open Country stratum.

| <u>Stratum</u> | <u>Total number of SUs</u> | <u>Sampling Rate</u> | <u>Number of SUs in Sample</u> |
|----------------|----------------------------|----------------------|--------------------------------|
| Urban | 4,100 | 1/100 | 41 |
| Rural Place | 500 | 1/100 | 5 |
| Open Country | <u>3,600</u> | 1/100 | <u>36</u> |
| Total | 8,200 | 1/100 | 82 |

The ultimate size of the sampling unit absorbs the small variation caused by rounding and does not introduce any bias in the procedure nor any change in the expectation in terms of total sample households. The expected number of 328 households may not be realized when the survey is actually conducted because of sampling error and change in the measure of size since 1950.

Further allocation of the sampling units to places and assignment of serial numbers to the sampling units are:

| <u>URBAN</u> | <u>No. of ODUs</u> | <u>Cumulative ODUs</u> | <u>Cumulative SUs</u> | <u>SU Serial Numbers Assigned</u> |
|---------------------|--------------------|------------------------|-----------------------|-----------------------------------|
| Asheville | 15,029 | 15,029 | 3738 | 0001 - 3738 |
| Urban Fringe | <u>1,455</u> | 16,484 | 4100 | 3739 - 4100 |
| Total | 16,484 | | | |
| <u>RURAL PLACE</u> | | | | |
| Biltmore Forest | 178 | 178 | 46 | 001 - 046 |
| Black Mountain | 407 | 585 | 150 | 047 - 150 |
| Jupiter | 37 | 622 | 160 | 151 - 160 |
| Enka | 487 | 1109 | 285 | 161 - 285 |
| Weaverville | 315 | 1424 | 366 | 286 - 366 |
| Swannanoa-Grovemont | <u>523</u> | 1947 | 500 | 367 - 500 |
| Total | 1947 | | | |

Using Enka as an example of the assignment of SUs to the Rural Places, the number of cumulative ODUs up to Enka is 622. This figure divided by the size of SU for Rural Place, 3.8940, is 159.7 or 160 when rounded to the nearest whole number. The number of cumulative ODUs through Enka is 1,109. The quotient, 1109 divided by 3.8940 is 284.8 or 285. The difference between 285 and 160, 125, is the number of sampling units assigned to Enka.

In the Open Country, the INOD total count is 12,784. The total number of sampling units assigned to the Open Country is 3600. Therefore, each sampling unit in the Open Country has an expected INOD size of 3.5511. In other words, for every 3.5511 INOD we expect about 4 ODUs. By following the procedure in the preceding paragraph, the number of SUs assigned to each division in the Open Country is as follows:

| <u>Division</u> | <u>INOD</u> | <u>Cumulative INOD</u> | <u>Cumulative No. of SUs</u> | <u>SU Serial Numbers Assigned</u> |
|-----------------|--------------|----------------------------|----------------------------------|---------------------------------------|
| 1 | 1,954 | 1,954 | 550 | 0001 - 0550 |
| 2 | 1,826 | 3,780 | 1064 | 0551 - 1064 |
| 3 | 1,479 | 5,259 | 1481 | 1065 - 1481 |
| 4 | 2,792 | 8,051 | 2267 | 1482 - 2267 |
| 5 | 1,514 | 9,565 | 2694 | 2268 - 2694 |
| 6 | 476 | 10,041 | 2828 | 2695 - 2828 |
| 7 | 1,470 | 11,511 | 3241 | 2829 - 3241 |
| 8 | <u>1,273</u> | 12,784 | 3600 | 3242 - 3600 |
| Total | 12,784 | | | |

C. The Sample Draw

With the assignment of sampling units, it is now possible to make the sample draw and to identify, within fairly broad

geographical limits, the location of the selected sampling units. It should be emphasized that the selection is based on sampling units, not on ODUs or on INOD. The selection of a particular sampling unit defines it uniquely, although at this stage it can be located only within a fairly broad area. Later, the identification process will be defined such that a particular unit can be located exactly.

The actual sample draw involves the use of random number tables to choose those sampling units which are to be designated as units in the sample. In the Urban stratum, 41 numbers between 1 and 4100 are selected at random and recorded in the order of draw. Similarly in the Rural Place stratum, five units are selected at random out of 500 and in the Open Country, 36 units are selected at random out of 3600.

The first two numbers selected in the Urban stratum are 2098 and 3800. Number 498 is the first selected in the Rural Place stratum and numbers 1269 and 1088 are the first two in the Open Country stratum. In the Urban stratum, SU #2098 falls in Asheville city, SU #3800 falls in the Urban Fringe, and is the 62nd sampling unit in the Urban Fringe (3800 - 3738, the number in Asheville, = 62). Similarly SU #498 in the Rural Place stratum is the 132nd unit located in Swannanoa-Grovement; in the Open Country, SU #1269 is the 205th sampling unit in Division 3 and SU #1088 is the 24th sampling unit in Division 3. To complete the identification of these sampling units, other materials are necessary:

1. The Urban Sample

The Block Statistics publication for Asheville is used to locate SU #2098. By following the procedure outlined above for

allocating sampling units to the various rural places, the census tract (ward) in which SU #2098 falls is identified.¹

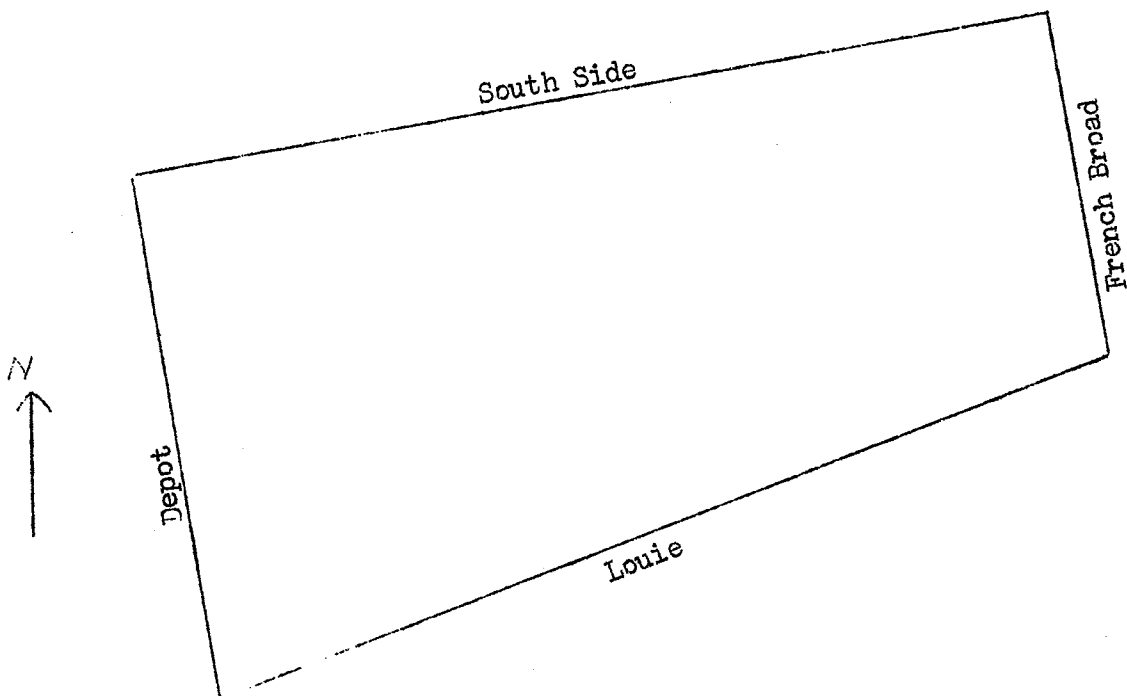
| <u>Ward</u> | <u>Number of ODUs</u> | <u>Cumulative ODUs</u> | <u>Cumulative SUs</u> | <u>SU Serial Numbers Assigned</u> |
|-------------|-----------------------|------------------------|-----------------------|-----------------------------------|
| 1 | 1166 | 1166 | 290 | 0001 - 0290 |
| 2 | 2927 | 4093 | 1018 | 0291 - 1018 |
| 3 | 2290 | 6383 | 1588 | 1019 - 1588 |
| 4 | 2584 | 8967 | 2230 | 1589 - 2230 |
| 5 | 2464 | 11431 | 2843 | 2231 - 2843 |
| 6 | 1447 | 12878 | 3203 | 2844 - 3203 |
| 7 | 494 | 13372 | 3326 | 3204 - 3326 |
| 8 | 352 | 13724 | 3414 | 3327 - 3414 |
| 9 | <u>1305</u> | 15029 | 3738 | 3415 - 3738 |
| Total | 15029 | | | |

SU #2098 is in Ward 4 - the 510th (2098-1588) SU of a total of 642 (2230-1588) in that ward. The number of ODUs by block is found in Table 3 - CHARACTERISTICS OF HOUSING FOR WARDS, BY BLOCKS: 1950.

¹From Table 2 - CHARACTERISTICS OF HOUSING BY WARDS: 1950; Asheville, N. C. Block Statistics, 1950 United States Census of Housing (H-E8) (The Asheville census tract boundaries coincide with the ward demarcations)

| <u>Ward</u> | <u>Block</u> | <u>Number of ODU's</u> | <u>Cumulative ODU's</u> | <u>Cumulative SUs</u> | <u>SU Serial Number Assigned (Within Ward)</u> |
|-------------|--------------|------------------------|-------------------------|-----------------------|--|
| 4 | 15 | 0 | | | |
| | 18 | 0 | | | |
| | 20 | 0 | | | |
| | 21 | 45 | 45 | 11 | 001 - 011 |
| | . | . | . | . | . |
| | . | . | . | . | . |
| | . | .. | . | . | . |
| | . | . | . | . | . |
| | . | . | . | . | . |
| | 278 | 48 | 2031 | 505 | . |
| | 279 | 26 | 2057 | 512 | 506 - 512 |

Thus, SU No. 2098 is the fifth (510 -505) SU of a total of 7 (512 - 505) in block No. 279, Ward 4. Block No. 279 is bounded by South Side, French Broad, Louie and Depot. A sketch of the block is given below.



The identification has now been completed down to a particular block. Seven sampling units have been assigned to this block and the particular one to be defined uniquely is the fifth among those seven. There are several different ways of determining this fifth sampling unit and the choice depends upon the particular situation and the information available. The methods fall into two main categories: segmenting and sub-sampling. Both of those words are, in a sense, misnomers, especially the second.

a. Segmenting

"Segmenting" is the division of the block into seven segments each with definable boundaries and each containing approximately equal numbers of ODUs.

Starting from some designated point and following some designated order, the sampling units (segments) are numbered from 1 through 7. For example, the segments could be numbered starting from the northeasternmost segment, proceeding in a clockwise direction. That one receiving the number 5 is the sampling unit selected for the sample.

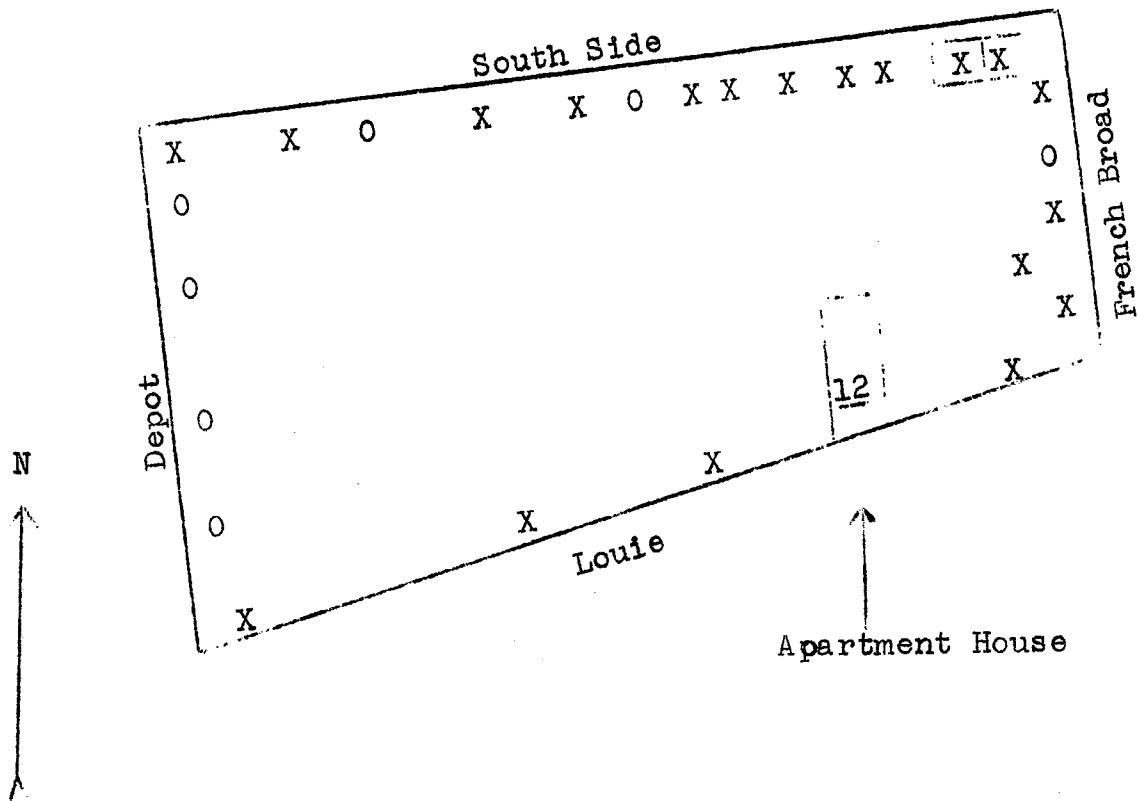
Segmenting may be accomplished in a number of different ways. Only three of the most common procedures are noted here.

- 1) If it is possible to cruise the area, the apparent ODUs can be spotted on a sketch. These, then, can be divided into seven segments of approximately equal size in terms of ODUs.
- 2) Sometimes an up-to-date map of the block is available showing the location of dwelling units. This could then be a basis for segmenting the block.
- 3) If a city directory is available, the block can be segmented on the basis of addresses around the block.

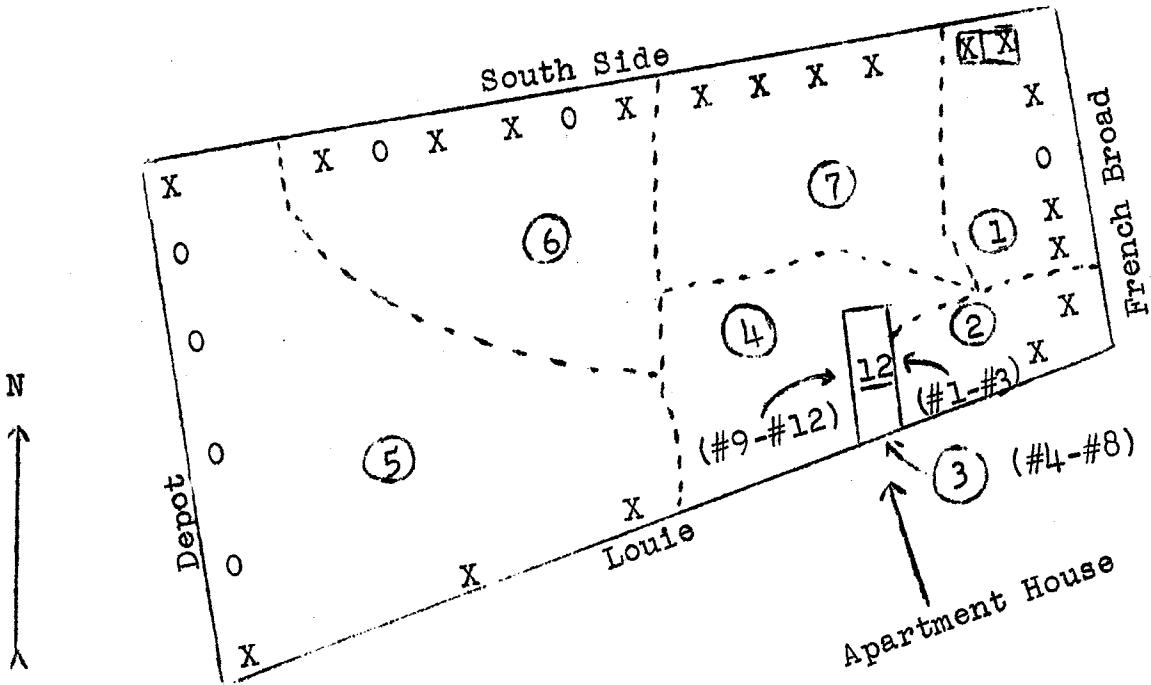
In both methods 1) and 2), care must be taken to insure that rules are made covering the assignment of all dwelling units in the

block to a unique segment. Assignment cannot be arbitrary, and, should not, of course, be left to the discretion of the interviewer.

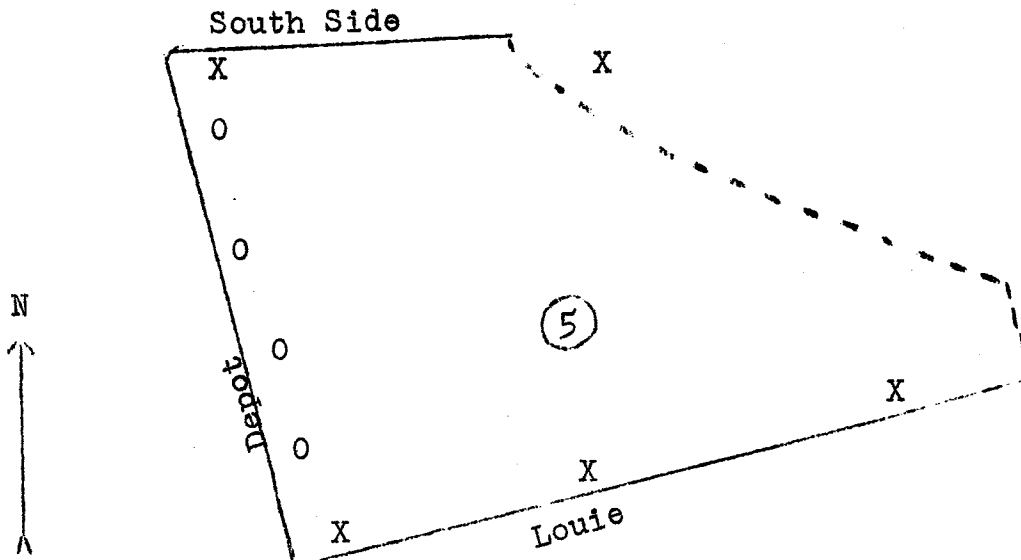
A cruise of the block might result in the following sketch information, where X indicates a possible ODU and O indicates a vacant dwelling unit:



Whereas the census information listed 26 ODUs, the current cruise information indicates 31 ODUs. Segmenting the block into 7 equal-sized segments would make segments of sizes 5, 5, 5, 4, 4, 4, 4 (an arbitrary but consistent rule could be made such that the first segments are the largest in the event that exactly equal sizes are impossible). Delineating the segments, which are now sampling units, and assigning sampling unit numbers make the sketch appear:



Sampling unit #5, which is in the sample starts at the first house west of the apartment house on Louie, goes to Depot, along Depot to South Side, and up to (but not including) the second house east on South Side. All occupied dwelling units, including



dwelling units thought to be vacant which turn out to be occupied and dwellings which are actually multiple units, are in the sampling unit so defined. Conversely, if the indicated ODU's are not occupied, less than 4 ODU's would be in the unit.

b. Sub-sampling

Sub-sampling is a misnomer in the sense that the sample has already been selected and the present procedure is merely identifying it. However, sub-sampling is the name in common usage.

Only two variations of this method are discussed here:

- 1) The entire block can be pre-listed by the interviewer and all the eligible households indicated. "Eligible" households are those households in which interviews are to be made according to the survey specifications. If all households are to be included, "all" households would be eligible. Starting from a pre-designated point (say, northeast corner of the segment) and moving in a pre-designated direction (either clockwise or counter-clockwise) every seventh eligible household is assigned for interview beginning with the fifth one that is eligible. (#5, #12, #19, #26, etc.) Had the second sampling unit been drawn, we would have started with the second eligible household and interviewed every seventh household (#2, #9, #16, #23, etc.).¹

¹ In cases where blocks contain a large number of sampling units (e.g., more than ten) the household identification is too laborious to be practical. To avoid this, a chunking procedure provides an unbiased means in selecting a final sampling unit. The large block is assigned "chunks", approximately equal in size, and the particular chunk to be segmented or sub-sampled in the field is marked appropriately. For example, a block containing 27 SUs of which the 22nd is in the sample might be chunked 10, 10, 7. The second SU of seven SUs in the third chunk is then, the 22nd SU in the block.

The field procedure is to cruise the block to obtain an estimate of the total number of ODU's and to locate the ODU's approximately on a sketch of the block. The quotient, estimated total number of ODU's divided by the number of SUs assigned to the block, is the "eye-size" of the sampling unit. Multiplying the eye-size by the number of SUs in each chunk gives the ODU numbers for each chunk. In this example, the estimated total number of ODU's is 77:

$$\frac{77}{27} = 2.9 \text{ (eye-size); } \begin{array}{l} \text{chunk 1 (10 SUs) = ODU's \#1-\#29, (10 x 2.9)} \\ \text{chunk 2 (10 SUs) = ODU's \#30-\#58} \\ \text{chunk 3 (7 SUs) = ODU's \#59-\#77} \end{array}$$

Chunk #3 is then dealt with according to the procedure specified for the smaller blocks.

- 2) In order to save the time and cost of pre-listing in a large area, the interviewer may be instructed to drive around the block to locate on a sketch what appears to be occupied dwelling units. These ODUs are then ordered, numbered and every seventh dwelling unit visited, beginning with the fifth. In this procedure, if a sample dwelling unit does not contain an eligible household no interview is taken and no substitution made. If two or more eligible households are found at what was thought to be one dwelling unit (it received one number), interviews are taken in all eligible households.

The materials used to locate SU #3800, which falls in the urban fringe, are the same as those used in identifying an SU in the Rural Place zone. The materials and procedure are described next.

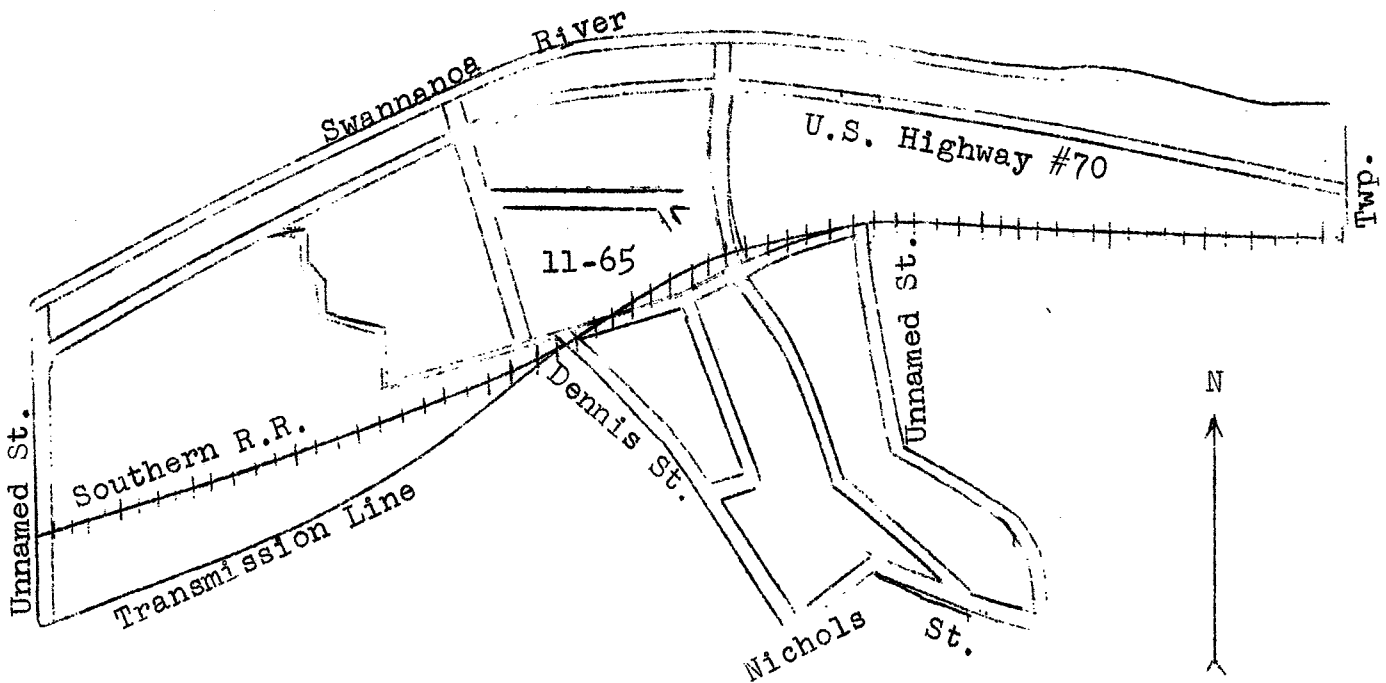
2. The Rural Place Sample

If Census ED maps and information have not previously been obtained, they must now be ordered for the urban fringe area and the rural places in which selected sampling units are located. Some time may be saved if aerial photographs of the same areas can be ordered at the same time. If eye estimates by cruising the places are to be made, aerial photographs will not be needed. The process will be demonstrated using SU #498 in the Rural Place stratum, which falls in Swannanoa-Grovemont, an unincorporated place.

EDs 11-64 and 11-65 comprise Swannanoa-Grovemont and have a total of 134 SUs assigned. The Census ED materials for this place list a total of 523 households, having heads, 14 years old and over; 211 in ED 11-64 and 312 in ED 11-65. These figures do not necessarily correspond exactly to the number of ODUs as it happened in this example, but the correspondence is close enough for the assignment of sampling units. There are an average of 3.9030 heads, 14 years old and over, per sampling unit. Hence, 54 SUs (#367-#420) are assigned to ED 11-64 and 80 SUs (#421-#500) are assigned to ED 11-65.

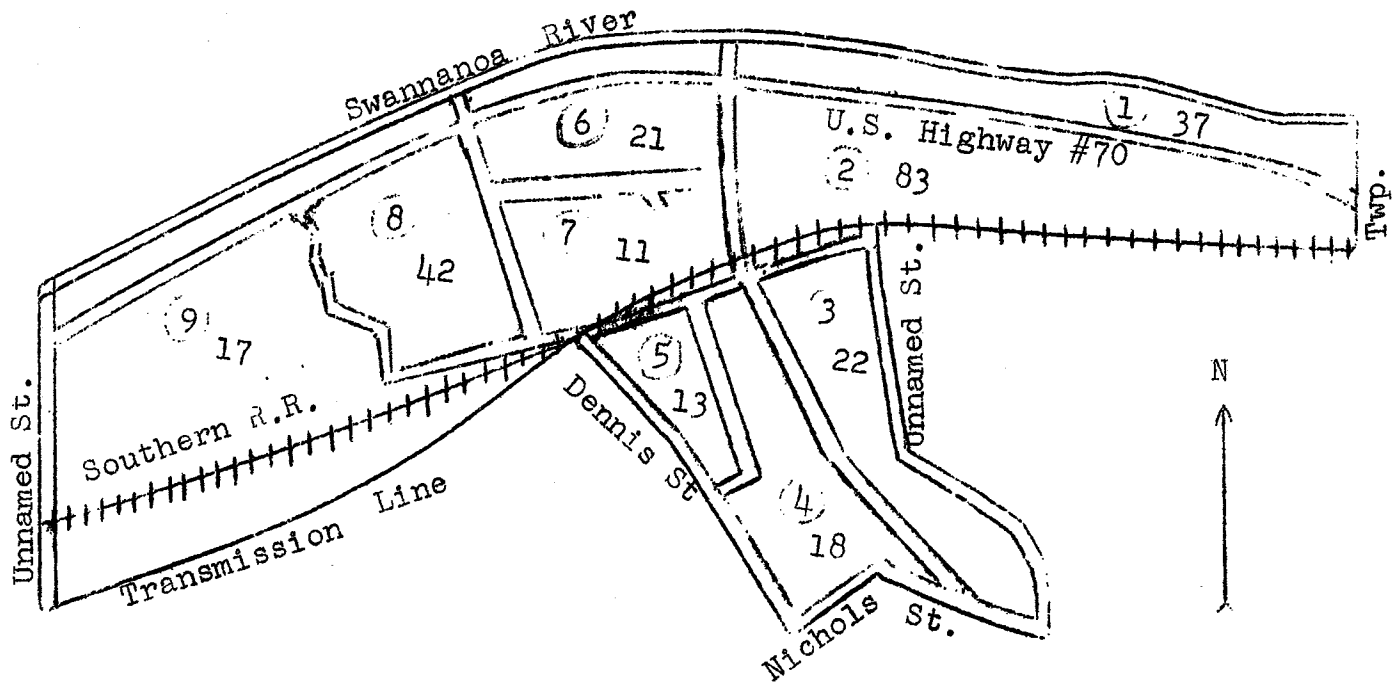
The sample SU is #498 or the 78th SU in ED 11-65. This is as far as the identification process can be taken without an aerial photograph or an estimate of ED 11-65 from a cruising operation. The following illustrations show the above process using aerial photo "ODU" counts:

A sketch of ED 11-65 is made from the Census ED map:



On an aerial photo of the area, ED 11-65 is delineated precisely:

.Numbering the blocks in the sketch and counting the houses in the aerial photo complete the sampling frame for ED 11-65.



| "Block" | Photo "ODUs" | Cumulative ODU's | Cumulative SUs | SU Serial Numbers Assigned (within ED 11-65) |
|---------|--------------|------------------|----------------|--|
| 1 | 37 | 37 | 11 | 01 - 11 |
| 2 | 83 | 120 | 36 | 12 - 36 |
| 3 | 22 | 142 | 43 | 37 - 43 |
| 4 | 18 | 160 | 48 | 44 - 48 |
| 5 | 13 | 173 | 52 | 49 - 52 |
| 6 | 21 | 194 | 59 | 53 - 59 |
| 7 | 11 | 205 | 62 | 60 - 62 |
| 8 | 42 | 247 | 75 | 63 - 75 |
| 9 | <u>17</u> | 264 | 80 | 76 - 80 |
| Total | 264 | | | |

$$\frac{264 \text{ (Photo ODU's)}}{80 \text{ (Assigned SUs)}} = 3.3 \text{ Photo ODU's/SU}$$

The ratio of the photo ODUs to the number of assigned sampling units is 3.3. By dividing the accumulated ODUs (the photo count) by 3.3, 80 sampling units are assigned to the "blocks". The 78th SU in ED 11-65 is the 3rd SU in block #9. The ultimate sampling unit can then be uniquely identified by segmenting or sub-sampling in the field. Re-examination of the photo should be made just in case segmenting could be accomplished in the office.

The above procedure is identical to that of the urban Block Statistics sampling process, except for the change in the measure of size. In this case, the change is from the Census information (3.8940) to the aerial photo measure (3.3). The expected size, in terms of 1950 Census ODUs, however, remains constant throughout the process, regardless of the measures of sizes used in subsequent phases.

3. The Open Country Sample

In the Open Country the procedure is analagous to the other strata. The new measure of size introduced here is the indicated number of dwellings (INOD) counted on the county highway map.

The first random number selected in the Open Country is 1269; the second, 1088. Both of these units are located in Division 3. The ratio of INOD to the number of sampling units assigned to the Open Country is 3.5511. The INOD counts are then accumulated in Division 3. Multiplying the random number by the INOD/SU ratio, 3.5511, indicates the location of the sampling units in the count units. The assignment of sampling units to the count units is done in the manner described in the Urban Rural Place examples:

| Random Number | x <u>3.5511</u> | From the INOD Listing | | Cumulative INOD | Cumulative No. of SUs | SU Serial Numbers Assigned |
|---------------|-----------------|-----------------------|------|-----------------|-----------------------|----------------------------|
| | | Count Unit | INOD | | | |
| 1269 | 4506 | 3-24 | 17 | 4488 | 1264 | |
| | | 3-25 | 22 | 4510 | 1270 | 1265-1270 |
| ----- | | | | | | |
| 1088 | 3864 | 3-2 | 25 | 3832 | 1079 | |
| | | 3-3 | 30 | 3862 | 1088 | 1080-1088 |
| | | 3-4 | 15 | 3877 | 1092 | 1089-1092 |
| ----- | | | | | | |

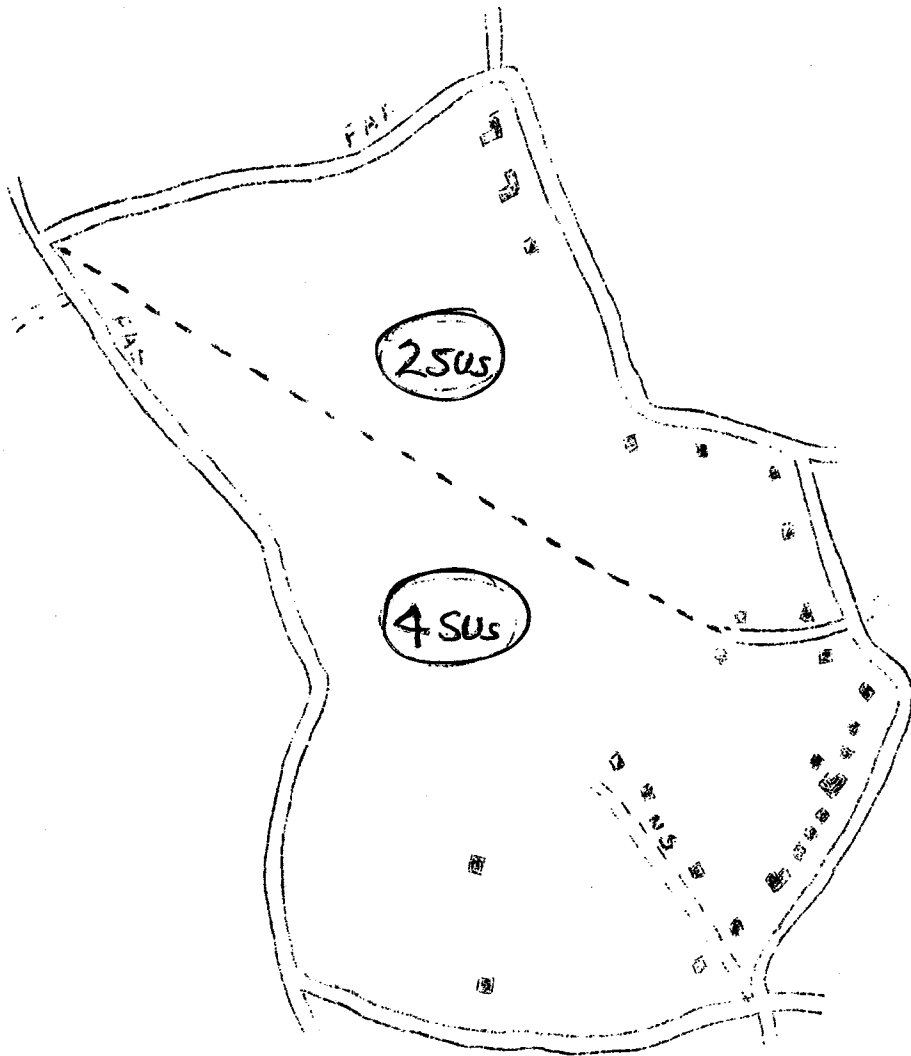
Sampling unit #1269 is, therefore, the 5th SU in count unit 3-25. Sampling unit #1088 is the 9th SU in count unit 3-3. Despite the fact that the product of the random number x 3.5511, 3864, fell in CU 3-4, the allocation of integral numbers of SUs to the count units is such that SU #1088 is in CU 3-3 and not, actually, in 3-4. This feature emphasizes the fact that sampling units are sampled, not the measure of size itself.

If feasible, the CUs should be segmented on the count map. This is accomplished by dividing the count unit along natural boundaries where possible. To avoid excessive costs of listing for sub-sampling, imaginary lines are sometimes used. Caution should be exercised in the use of imaginary lines, however, for the sampling unit must be definable and identifiable in the field. In most cases, anchoring the end-points of the imaginary lines to road intersections, road extensions, bridges, churches, etc. makes the imaginary line easily identified in the field.

Aerial photos are excellent for segmenting the Open Country count units. Farm roads and lanes, field boundaries, tree lines, etc. provide good boundaries which are easily identified in the

field. The disadvantages of aerial photo use are the time involved in picking and ordering photos and the cost of the photos.

If segmenting in the office is impractical, either segmenting or sub-sampling in the field must be done. In some cases, the count unit can be partially segmented; for example, CU 3-25 might be segmented into 2 parts, 4 SUs in one, 2 SUs in the other. The 5th SU then would be the first in the segment with 2 SUs, and finally identified in the field by further segmenting or sub-sampling.



IV Some Standard Sample Designs

A. Simple Random Sampling

A simple random sample may be defined as one in which n units are selected out of N so that each of the $\binom{N}{n}$ possible samples has an equal chance for selection. As indicated earlier, an estimate of the mean per sampling unit is of little interest but estimates of the population total or estimates of the mean per element are often desired. These estimates, and estimates of their variance are:

The estimated population total (T):

$$T = \frac{N}{n} \sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij} \quad (1)$$

where T = the estimated population total

N = number of sampling units in the universe

n = number of sampling units in the sample

$\frac{N}{n}$ = reciprocal of the sampling rate

m_i = number of elements in the i-th sampling unit

y_{ij} = measurement on the j-th element in the i-th sampling unit.

The estimated variance of the population total (s_T^2):

$$s_T^2 = N(N-n) s^2/n \quad (2)$$

where

$$s_y^2 = \sum_{i=1}^n \left(\sum_{j=1}^{m_i} y_{ij} - \bar{y}_u \right)^2 / (n-1)$$

$$\bar{y}_u = \sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij} / n = \text{mean per sampling unit.}$$

The estimated mean per element (\bar{y}):

$$\bar{y} = \sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij} / \sum_{i=1}^n m_i \quad (3)$$

The estimated variance of the mean per element ($s_{\bar{y}}^2$) is approximated by:

$$s_{\bar{y}}^2 = \frac{N-n}{nN} \bar{y}^2 \left\{ \frac{s_y^2}{\bar{y}^2} + \frac{s_m^2}{\bar{m}^2} - \frac{2s_{ym}}{\bar{y}\bar{m}} \right\} \quad (4)$$

where $s_m^2 = \sum_{i=1}^n (m_i - \bar{m})^2 / (n-1)$

$$\bar{m} = \sum_{i=1}^n m_i / n$$

$$s_{ym} = \sum_{i=1}^n \left(\sum_{j=1}^{m_i} y_{ij} - \bar{y}_u \right) (m_i - \bar{m}) / (n-1).$$

The estimate of the population total (T) is an unbiased estimate of the true population total and its estimated variance is unbiased. However, the estimate \bar{y} is in reality a ratio estimate since both the numerator and the denominator are random variables and is subject to a bias. This bias is usually considered to be negligible in large samples. The expectation of the estimated variance is only an approximation to the true variance of \bar{y} , but again it is considered to be a satisfactory approximation in large samples.

Simple random area sampling is ordinarily used only when the universe is confined to one of the three zones, i.e. urban, rural place or open country. For example, in sampling for agricultural items, the universe may be limited to the open country zone.

B. Stratified Random Sampling

For general household or population surveys in which the universe includes all three zones, it is natural to stratify by zone since Census information is readily available by zone. The statistical advantage in using stratified random sampling over simple random sampling is realized when homogeneous strata are constructed so that differences between strata are removed from the sampling error, thus increasing efficiency. Since characteristics of elements may differ by zone we can usually expect a moderate gain in efficiency by zonal stratification.

A stratified random sample may be described most simply as merely a simple random sample within each of k strata.

1. General Case - Unequal Sized Strata

Estimated totals and means per element with corresponding estimated variances are:

The estimated population total (T):

$$T = \sum_{h=1}^k N_h \bar{y}_{uh} \tag{5}$$

$$= \frac{N}{n} \sum_{h=1}^k \sum_{i=1}^{n_h} \sum_{j=1}^{m_i} y_{hij} \tag{6}$$

where N_h = number of sampling units in the population in the h -th stratum

k = number of strata

$$\bar{y}_{uh} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_i} y_{hij} / n_h$$

n_h = number of sampling units in the sample in the h -th stratum

$$n = \text{number of SU in the entire sample} = \sum_{h=1}^k n_h .$$

The estimated variance of the population total (s_T^2):

$$s_T^2 = \sum_{h=1}^k N_h (N_h - n_h) s_h^2 / n_h \tag{7}$$

where $s_h^2 = \sum_{i=1}^{n_h} \left(\sum_{j=1}^{m_i} y_{hij} - \bar{y}_{uh} \right) / (n_h - 1).$

The estimated variance of the population total (s_T^2) when the within stratum variance are equal:

$$s_T^2 = N(N-n) s_w^2 / n \tag{8}$$

where $s_w^2 = \sum_{h=1}^k \sum_{i=1}^{n_h} \left(\sum_{j=1}^{m_i} y_{hij} - \bar{y}_{uh} \right)^2 / (n-k).$

The mean per element (\bar{y}):

$$\bar{y} = \frac{\sum_{h=1}^k N_h \bar{y}_{uh}}{\sum_{h=1}^k N_h \bar{m}_h} \quad (9)$$

$$= \frac{\sum_{h=1}^k \sum_{i=1}^{n_h} \sum_{j=1}^{m_i} y_{hij}}{\sum_{h=1}^k \sum_{i=1}^{n_h} m_i} \quad (10)$$

where $\bar{m}_h = \sum_{i=1}^{n_h} m_{hi} / n_h$.

The estimated variance of this mean per element is approximated by:

$$s_{\bar{y}}^2 = \bar{y}^2 \left(\frac{n-n_h}{Nn} \right) \left(\frac{s_{wy}^2}{\bar{y}_u^2} + \frac{s_{wm}^2}{\bar{m}^2} - \frac{2s_{wym}}{\bar{y}_u \bar{m}} \right) \quad (11)$$

where s_{wy}^2 is the within stratum mean square for y

s_{wm}^2 is the within stratum mean square for m

$$\bar{y}_u = \sum_{h=1}^k N_h \bar{y}_{uh} / N$$

$$\bar{m} = \sum_{h=1}^k N_h \bar{m}_h / N.$$

The estimate of \bar{y} , given in alternative forms in (9) and (10), is known as the combined stratum ratio estimate. See Hansen, Hurwitz and Gurney (1946). This estimate is also subject to bias but it is not likely that the bias will be serious.

It is possible, of course, to use other estimators. If the means per element differ widely from stratum to stratum, and the sample size is sufficiently large,

a separate stratum ratio estimate should be considered. See Cochran (1953).

2. Special Case - Equal Sized Strata

It has been shown that only modest gains in efficiency are realized by further geographic stratification. See Jessen and Houseman (1944). However, when the size of sample is sufficiently large so that the degrees of freedom are adequate for reasonable precision in the estimate of the sampling error, it is common to "stratify to the hilt". For this procedure, substrata of equal size in terms of sampling units are formed within each stratum or zone. Two sampling units are selected at random from each stratum to provide an estimate of variance. Thus any proximity correlation of SUs will result in a gain, though this gain may be small. In addition, this method of stratification insures a reasonable scatter of the sampling units. Little additional office work is required to draw a sample of two from each of $n/2$ strata. The utilization of this design, however, may require another adjustment in the allocation of sampling units to the various zones, since it is now necessary to have not only an integral number of sampling units, in the sample from each stratum (zone), but this integer must be even. For example, the new adjusted allocation for the three zones in Buncombe county, might be as follows:

| <u>Stratum</u> | <u>No. of ODU's</u> | <u>No. of SUs</u> | <u>Size of SU</u> | <u>No. of Sub-Strata</u> |
|----------------|---------------------|-------------------|-------------------|--------------------------|
| Urban | 16,484 | 4,200 | 3.9248 | 21 |
| Rural Place | 1,947 | 400 | 4.8675 | 2 |
| Open Country | 14,418 | 3,600 | 4.0050 | 18 |
| Total | 32,849 | 8,200 | 4.0050 | 41 |

The average size of sampling unit is still close to four for both the Urban and Open Country zones but is now almost five for the Rural Place stratum. Since the population is relatively large in both the Urban and Open Country there is little effect of the adjusted numbers of SU on the size of the SU. However, with the small population in the Rural Place stratum an adjustment in the number of SUs allocated to that stratum is reflected to a greater extent in the size of the SU in that stratum. The differences in cluster sizes from stratum to stratum is no cause for concern.

The breakdown of the Urban, Rural Place and Open Country allocation is now carried through as before using the new allocations and the new SU sizes. The sample draw is made in about the same manner except that one small additional step is required. Since there are 41 sub-strata there are exactly 200 sampling units per sub stratum. Hence, we draw 41 sets of two numbers each at random from 001-200; 21 sets will apply to the Urban area, two to the Rural Place area and 18 to the Open Country. The draw is without replacement within a set, i.e. no two numbers may be the same within a given set. However, they can be repeated in a different set. The procedure will be illustrated with the open country stratum.

| Stratum | Sub-stratum | Random No. in Sub-Stratum | Random No. in Stratum |
|--------------|-------------|---------------------------|-----------------------|
| Open Country | 1 | 9 | 9 |
| | | 51 | 51 |
| | 2 | 127 | 327 |
| | | 30 | 230 |
| | 3 | 8 | 408 |
| | | 192 | 592 |
| | . | . | . |
| | . | . | . |
| | . | . | . |
| | 18 | 122 | 3522 |
| 200 | | 3600 | |

To obtain the random numbers within the stratum for sub-stratum 2, we take the numbers selected (between 1 and 200) and add 200 to them. The two random numbers selected in sub-stratum 2 were 127 and 30. Adding 200 gives the 327th SU in the open country and the 230th. Similarly in sub-stratum 3 we add 400 to the random draw and in 18, we add 3400. After the random numbers in the stratum have been calculated, the procedure for locating the selected SUs is exactly the same as that described previously.

The estimates and estimated variances can be computed by (5), (6), (7) and (8). A short-cut method of computing s_w^2 can be employed as follows when only two sampling units are selected from each stratum:

$$s_w^2 = \sum_{h=1}^{n/2} \frac{\left(\sum_{j=1}^{n/2} y_{h1j} - \sum_{j=1}^{m_2} y_{h2j} \right)^2}{n} \quad (12)$$

where h now denotes the particular sub-stratum. In other words, the square of the difference between the SU totals summed over all $n/2$ strata and divided by n will give an estimate of s_w^2 . The estimate of s_w^2 can then be inserted into formula (8).

3. Optimum versus Proportional Allocation.

There are two principal procedures for allocating the sample to the various strata. These are commonly known as proportional allocation and optimum allocation. An allocation is proportional when the number of sampling units in the sample assigned to a given stratum is proportional to the number of sampling units in that stratum in the universe, i.e. $n_h \propto N_h$.

Optimum allocation requires the assumption of a given cost function. If the cost of enumerating in a given sampling unit varies from stratum to stratum, a simple cost function of the form

$$c = \sum_{h=1}^k c_h n_h \quad (13)$$

is reasonable. If such a cost function is assumed, the allocation is optimum if the number of sampling units in the sample assigned to a given stratum is proportional to the product of the number of SUs in the universe in that stratum and the stratum standard deviation divided by the square root of the cost of enumerating in an SU in the prescribed stratum, i.e. $n_h \propto N_h \sigma_h / \sqrt{c_h}$.

Throughout this discussion the use of proportional allocation has been advocated for a number of reasons; in fact formulae (6), (8) and (10) are not applicable unless sampling is proportional. These reasons are:

- (i) Usually estimates from sample data are made for a large number of items (populations) for any given universe. Each of these populations may have a different standard error thus requiring a different optimum allocation of the sample. Since this is obviously impossible for a given sample size, some compromise is necessary. Often proportional allocation is a reasonable compromise.
- (ii) Even for special purpose surveys involving only one population, estimates of cost and variability must be available. Cochran (1953) states "as a working rule, proportional allocation is usually to be recommended unless the expected gain in precision from optimum allocation, as estimated in advance of taking the sample, exceeds 20 per cent."
- (iii) The sample draw is easier to make and estimates are simpler to compute. For example compare formula (6) with (5) and (8) with (7). A stratified random sample with proportional allocation is sometimes known as a "self-weighting sample".

U. Two-Stage Sampling

In order to reduce costs and/or to take advantage of the availability of field crews at certain branch locations, sampling is sometimes carried on in stages. That is the universe will be divided into fairly large geographical areas denoted as first-stage units (FSUs). Perhaps the most commonly used first-stage unit is the county. Within each FSU, second-stage units (SSUs) are constructed. In practice one or more FSU are selected in accordance with a prescribed procedure and within these sample FSU, one or more sample SSUs are selected. The SSUs correspond to the area sampling units used in earlier discussion. A two-stage sampling system has the characteristic of concentrating the sample around several "sample points" rather than spreading it over the entire universe. In general, the two-stage sample will be less precise than a stratified random sample of the same size (in terms of SSUs) but if the cost is considerably less, the two-stage system may be more efficient.

1. Selection of the First-Stage Units with Unequal Probabilities

a. One Sample FSU per Stratum

At this writing the sampling system enjoying most favor in large scale surveys is one first advocated by Hansen and Hurwitz (1943) and is designed as follows:

- (i) Determine the number of FSUs or sampling points desired. This decision may be based upon the field organization or perhaps the cost. Rarely is information available on which to make an optimum choice.
- (ii) Construct strata equal in number to the FSUs where the FSUs comprising each stratum are contiguous. A measure of size, which we hope is highly correlated with the principal items of interest, is attached to each FSU. Then one FSU is selected at random from each stratum with probability proportional to this measure of size.
- (iii) From the overall sampling rate desired, the number of SSUs required in each selected FSU is computed as follows:

$$P_{hr} = \frac{gP_{hr}}{Z_{hr}} \quad (14)$$

where P_{hr} = the universe number of SSUs in the r^{th} FSU in the h^{th} stratum

p_{hr} = the sample number of SSUs in the r^{th} FSU in the h^{th} stratum

g = the overall sampling fraction in all strata

Z_{hr} = the probability of selecting the r^{th} FSU, where the

Z_{hr} summed over r equals 1 in each stratum.

Note two points.

- (i) We need to know P_{hr} for the selected FSU only.
- (ii) If $Z_{hr} = \frac{P_{hr}}{P_h}$

where P_h is the universe number of SSUs in the h^{th} stratum.

Then

$$P_{hr} = gP_h \quad (15)$$

In this situation, no matter which of the FSU are selected in the h^{th} stratum, the number of SSUs in the sample will be the same.

Further, if all P_h are equal, i.e. we construct equal sized strata, the sample size will be the same in each stratum.

The estimated population total (T):

$$T = \frac{1}{g} \sum_{h=1}^k P_{hr} \sum_{i=1}^{m_{hri}} \sum_{j=1}^{m_{hri}} y_{hij} \quad (16)$$

which is simply the sample total multiplied by the reciprocal of the overall sampling fraction. No unbiased estimate of the variance of T is possible from sample data.

The estimated mean per element (\bar{y}):

$$\bar{y} = \frac{\sum_{h=1}^k P_{hr} \sum_{i=1}^{m_{hri}} \sum_{j=1}^{m_{hri}} y_{hij}}{\sum_{h=1}^k P_{hr} \sum_{i=1}^{m_{hri}} m_{hi}} \quad (17)$$

which is the sample total for y divided by the number of elements in the sample. Again no unbiased estimate of variance is possible.

b. Two or More Sample FSU per Stratum

Those who favor selecting one FSU per stratum argue that stratification is desperately needed and state that they are willing to sacrifice unbiased estimates of variance in order to receive a gain in precision. As indicated above, the use of such a system does not permit its internal evaluation; hence, the validity of that argument cannot be ascertained from sample data. There are many, however, who are not satisfied with biased estimates of error. Further, there is some evidence that geographic stratification results only in moderate gains in

precision. Therefore, considerable work has been done on developing sampling systems which will take advantage of the measures of size in assigning unequal probabilities but will allow an unbiased estimate of the sampling error. The recent literature includes Horvitz and Thompson (1952), Yates and Grundy (1953), Sen (1953) and Des Raj (1956) who cite a number of additional references.

A wide range of selection probabilities and various estimators for each have generated a large number of sampling systems for which unbiased estimates of variance can be obtained from sample data. Adequate treatment of this subject is beyond the scope of this handbook. The sample draw and the estimate of the total will be illustrated below for one such system falling in this category. For further details, including variance estimators, see the references quoted immediately above.

In this design two FSUs are selected without replacement within each stratum; the first at random with probability proportional to some measure of size, X_{hr} , where

$$X_{hr} = \sum_{i=1}^{P_{hr}} \sum_{j=1}^{M_{hri}} x_{hrij}, \quad (18)$$

and the second at random from those remaining with equal probability. This is equivalent to forming all possible pairs within a given stratum, obtaining a sum of the sizes for each pair, cumulating these sums and finally selecting one pair with probability proportional to its sum. The SSUs are to be selected with equal probability and without replacement. Then

$$P_{hr} = \frac{g_{hr}^P}{Z'_{hr}} \quad (19)$$

where

P_{hr} , g and P_{hr} are defined as before and Z'_{hr} is the a priori probability of including the r^{th} FSU in the sample of two FSU. This probability is sometimes referred to as inclusion probability as contrasted with

selection probability. For example, suppose we have three FSUs in a given stratum with the following values for the auxiliary variable:

| <u>FSU</u> | <u>X_{hr}</u> |
|------------|-----------------------|
| 1 | 7 |
| 2 | 2 |
| <u>3</u> | <u>1</u> |
| Sum | 10 |

$$\text{Then } Z'_{hl} = (7/10)(1/2) + (2/10)(1/2) + (7/10)(1/2) + (1/10)(1/2) = 17/20$$

An unbiased estimate of the population total (T):

$$T = \frac{1}{g} \sum_{h=1}^k \sum_{r=1}^{n_h} \sum_{i=1}^{p_{hr}} \sum_{j=1}^{m_{hri}} y_{hrij} \quad (20)$$

where $n = 2 =$ the number of FSU selected per stratum.

The estimated mean per element (\bar{y}):

$$\bar{y} = \frac{\sum_{h=1}^k \sum_{r=1}^{n_h} \sum_{i=1}^{p_{hr}} \sum_{j=1}^{m_{hri}} y_{hrij}}{\sum_{h=1}^k \sum_{r=1}^{n_h} \sum_{j=1}^{m_{hri}} m_{hri}} \quad (21)$$

2. Selection of the First-Stage Units with Equal Probability

It is possible, in an entirely different manner, to take advantage of the measures of size which may be available for counties, townships or other political or geographic subdivisions. As indicated earlier, two-stage sampling is popular in large scale sample surveys because the SSUs are clustered at sample points thus reducing the cost and making use of any existing field staff. It is possible to create FSUs of equal size, in terms of SSUs so that the objectives of two-stage sampling are achieved, yet the design becomes a simple "nested" sample. The construction of the FSUs is accomplished in exactly the same manner as the construction of the equal sized sub-strata described in IV-B-2.

The sample FSUs are selected at random from the universe (or within a stratum) without replacement and with equal probability. This is a special case of sampling with probabilities proportional to size; here all FSUs are of equal size. Within each sample FSU, the same number of SSUs are selected at random without replacement and with equal probability. The SSUs are then identified on the map in the usual manner. The SSUs, and hence the observational units, will be clustered together in much the same manner as for the selection of the FSUs with unequal probability. However, instead of being contained entirely within a county, SSUs within the same FSU may be found in parts of adjacent counties since political boundary lines are ignored when the FSUs are constructed. This increases the initial map work in the office prior to the field work and adds a little expense for extra maps and attendant materials; however, the analysis -- particularly the estimation of the sampling error -- is much simpler. An unbiased estimate of the total assuming stratification with proportional allocation and two FSU per stratum is exactly the same as (20), i.e. the estimated total in the population is the product of the reciprocal of the sampling rate and the sample total. Similarly, the estimate of the mean per observational unit (\bar{y}) is the same as given in (21). Both n_h and p_{hr} will be constants under this system, and can be replaced by the appropriate constant when decided upon.

The estimated variance of the total (s_T^2) can be taken directly from the analysis of variance.

| <u>Source of Variation</u> | <u>Degrees of Freedom</u> | <u>Mean Square</u> |
|----------------------------|-----------------------------|--------------------|
| Between Strata | $k - 1$ | |
| Between FSU in Strata | $k(n-1)$ | B |
| <u>Between SSU in FSU</u> | <u>$kn(p-1)$</u> | W |
| Total | $knp - 1$ | |

Then

$$s_T^2 = \frac{(NP)^2 B}{np} \quad (22)$$

where N = number of FSUs in the universe

P = number of SSUs per FSU in universe.

If n/N , (the sampling rate of FSUs) exceeds .05, the finite population correction factor should be applied, in which case (22) becomes

$$s_T^2 = \frac{(NP)^2}{np} \left[\left(\frac{N-n}{N}\right)B + \left(\frac{P-p}{P}\right)\frac{n}{N}W \right] \quad (23)$$

In concluding this section on the use of the materials, treatment of some sampling systems has been necessarily inadequate for the researcher to apply without some knowledge of sampling. No attempt has been made to prove the results shown here, nor in many cases, to explain the rationale behind the use of a particular design. Therefore, as in the case with any "cook book" type of exposition, caution must be exercised in the selection of the ingredients. If a problem has unusual features which may complicate the design, the advice of a trained sampling statistician may prove worthwhile.

References

- Cochran, W. G. (1953). *Sampling Survey Techniques*, John Wiley and Sons.
- Des Raj (1956). "Some estimators in sampling with varying probabilities without replacement." *Jour. Amer. Stat. Assoc.* 51:269-284.
- 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 Gurney, M. (1946). "Problems and methods of a sample survey of business." *Jour. Amer. Stat. Assoc.* 41:173-189.
- Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory*, John Wiley and Sons.
- Horvitz, D. G. and Thompson, D. J. (1952). "A generalization of sampling without replacement from a finite universe." *Jour. Amer. Stat. Assoc.* 47:663-684.
- Houseman, E. E. and Reed, T. J. (1954). "Application of probability area sampling to farm surveys." *Agri. Handbook No. 67*, U. S. Department of Agriculture, Agri. Marketing Service.
- Jessen, R. J. (1947). "The Master Sample Project and its use in agricultural economics." *Jour. Farm Econ.* 29:531-540.
- King, A. J. and Jessen, R. J. (1945). "The Master Sample of Agriculture." *Jour. Amer. Stat. Assoc.* 40:38-56.
- Sen, A. R. (1953). "On the estimate of the variance in sampling with varying probabilities." *Jour. Indian Soc. Agri. Stat.* Vol V, No. 2. 119-127.
- Yates, F. and Grundy, P. M. (1953). "Selection without replacement from within strata with probability proportional to size." *Jour. Roy. Stat. Soc. Series B*, 15, No. 1, 253-261.

APPENDIX

A. CENSUS DEFINITIONS

1. Urban and Rural Residence - Urban housing comprises all dwelling units in (a) places of 2,500 inhabitants or more incorporated as cities, boroughs, and villages, (b) incorporated towns of 2,500 inhabitants or more except in New England, New York and Wisconsin, where "towns" are simply minor civil divisions of counties, (c) the densely settled urban fringe around cities of 50,000 inhabitants or more, including both incorporated and unincorporated areas, and (d) unincorporated places of 2,500 inhabitants or more outside any urban fringe. The remaining dwelling units are classified as rural.

The rural classification comprises a variety of residences, such as isolated homes in the open country, dwelling units in villages and hamlets of fewer than 2,500 inhabitants, and some dwelling units in the areas surrounding urban places of fewer than 50,000 inhabitants.

2. Farm and Non-Farm Residence - In the 1950 Census, the enumerators in rural areas were specifically instructed to base the farm-non-farm classification of a dwelling unit on the respondent's answer to the question, "Is this house on a farm?" Farm residence is therefore determined without regard to the occupation of the members of the household. The classification depends upon the respondent's conception of what is meant by the word "farm", and consequently reflects local usage rather than the uniform application of an objective definition. For this reason, there is considerable variability of response among families living in areas where farm operation is part-time or incidental to other activities.

Dwelling units located on farm land for which cash rent was paid for the house and yard only, and dwelling units on institutional grounds and in summer camps and tourist courts, were classified as non-farm, regardless of the answer to the above question.

In rural areas, dwelling units are classified into rural-farm units which comprise all dwelling units on farms, and rural-non-farm units which are the remaining rural units. In most areas, virtually all farm housing is in rural areas. Therefore, housing characteristics are shown in this report for rural-farm dwelling units instead of for all farm units. Urban-farm dwelling units are dwelling units on farms within the boundaries of urban areas. Such farms constitute only 1.5 percent of the total farm dwelling units in the United States.

3. Types of Places - The term "place" refers to a concentration of population regardless of legally prescribed limits, powers, or functions. Thus, some areas having the legal powers and functions characteristic of incorporated places are not recognized as places.

In a majority of instances, however, the legally prescribed limits of incorporated places serve to define concentrations of population.

In addition to incorporated places, the 1950 Census recognizes 1,430 unincorporated places. These unincorporated places, which contain heavy concentrations of population, are recognized as places by virtue of their physical resemblance to incorporated places of similar size. To make this recognition possible, the Bureau of the Census has defined boundaries for all unincorporated places of 1,000 inhabitants or more which lie outside the urban fringes of cities of 50,000 inhabitants or more. Because local practice as to incorporation varies considerably from one part of the country to another, some States have very few if any such unincorporated places and others have a great many. Although there are also unincorporated places within the urban fringe, it was not feasible to establish boundaries for such places, and, therefore, they are not separately identified.

Urban places are incorporated places which had 2,500 inhabitants or more and those unincorporated places outside of urban fringes which had 2,500 inhabitants or more on April 1, 1950. In many states, the number of dwelling units in "urban places" is somewhat less than the total urban. The difference comprises dwelling units in those portions of urban fringes that are incorporated places of less than 2,500 inhabitants or are unincorporated.

Places of 1,000 to 2,500 inhabitants comprise incorporated places and those unincorporated places outside urban fringes which had from 1,000 to 2,500 inhabitants on April 1, 1950. In 1940, places of this size for which data were presented were limited to incorporated places.

4. Standard Metropolitan Areas - Except in New England, a standard metropolitan area is a county or group of contiguous counties which contains at least one city of 50,000 inhabitants or more. Counties contiguous to the one containing such a city are included in a standard metropolitan area if according to certain criteria they are essentially metropolitan in character and socially and economically integrated with the central city. Standard metropolitan areas are not confined within state boundaries nor within region or division boundaries. For states having standard metropolitan areas, the constituent counties are found in tables 17 to 21 in the state chapters.

5. Urbanized Areas - Each urbanized area contains at least one city with 50,000 inhabitants or more in 1940 or according to a special census between 1940 and 1950. Each urbanized area also includes the surrounding closely settled incorporated places and unincorporated areas that comprise its "urban fringe." The boundaries of these fringe areas were established to conform as nearly as possible to the actual boundaries of thickly settled territory, usually characterized by a closely spaced street pattern. Like standard metropolitan areas, urbanized areas are not confined within state boundaries, nor within region or division boundaries. A complete description and a map of each urbanized area of a state is at the end of the state chapter.

The urbanized area can be characterized as the physical city as distinguished from both the legal city and the metropolitan community. In general, the urbanized area represents the thickly settled urban core of the standard metropolitan area. Urbanized areas are smaller than standard metropolitan areas and in most cases are contained in them. Since the boundaries of standard metropolitan areas are

determined by county lines and those of urbanized areas by the pattern of urban growth, there are small segments of urbanized areas, in a few instances, which lie outside the standard metropolitan area.

6. Dwelling Unit - In general, a dwelling unit is a group of rooms or a single room occupied or intended for occupancy as separate living quarters by a family or other group of persons living together or by a person living alone.

Ordinarily, a dwelling unit is a house, an apartment, or a flat. A dwelling unit may be located in a structure devoted to business or other nonresidential use, such as quarters in a warehouse where the watchman lives, or a merchant's quarters in back of his shop. Trailers, boats, tents, and railroad cars, when occupied as living quarters, are included in the dwelling unit inventory.

A group of rooms, occupied or intended for occupancy as separate living quarters, is a dwelling unit if it has separate cooking equipment or a separate entrance. A single room, occupied or intended for occupancy as separate living quarters, is a dwelling unit if it has separate cooking equipment or if it constitutes the only living quarters in the structure. Each apartment in a regular apartment house is a dwelling unit even though it may not have separate cooking equipment. Apartments in residential hotels are dwelling units if they have separate cooking equipment or consist of two rooms or more.

Living quarters of the following types are not included in the dwelling unit inventory: rooming houses with five lodgers or more, transient accommodations (tourist courts, hotels, etc., predominantly for transients), and barracks for workers (railroad, construction, etc.). Living quarters in institutions (for delinquent or dependent children, for handicapped persons, for the aged, for prisoners, etc.), general hospitals, and military installations are likewise excluded from the dwelling unit inventory except for dwelling units in buildings containing only family quarters for staff members.

7. Occupancy Characteristics -

(a) Occupied dwelling unit - A dwelling unit is occupied if a person or group of persons was living in it at the time of enumeration or if the occupants were only temporarily absent, as for example, on vacation. However, a dwelling unit occupied at the time of enumeration by nonresidents is not classified as occupied but as a "nonresident" dwelling unit.

Household - A household consists of those persons who live in a dwelling unit; by definition, therefore, the count of occupied dwelling units is the same as the count of households. However, there may be small differences between these counts in the Housing and the Population reports because the data were processed independently.

(b) Nonresident dwelling unit - A nonresident dwelling unit is a unit which is occupied temporarily by persons who usually live elsewhere. Nonresident units are not included with occupied dwelling units.

(c) Vacant dwelling unit - A dwelling unit is vacant if no persons were living in it at the time of enumeration, except when its occupants were only temporarily absent.

B. HOUSING DATA SOURCES

Housing data are compiled for the following geographical and political divisions.

1. Continental United States
2. Regions and divisions
3. States

4. State Economic Areas - State economic areas are relatively homogeneous subdivisions of states. They consist of single counties or groups of counties which have similar economic and social characteristics. The boundaries of these areas have been drawn in such a way that each state is subdivided into a few parts, with each part having certain significant characteristics which distinguish it from the other areas which it adjoins. The 48 states have been subdivided into 501 state economic areas. In any one application, the number of areas is substantially reduced by combining some areas.

5. Standard Metropolitan Areas

6. Counties

7. Urbanized Areas

8. Urban Places

9. Rural Places of 1,000 to 2,500 population

10. Census Tracts - Census tracts are small areas, having a population usually between 3,000 and 6,000, into which certain large cities (and sometimes their adjacent areas) have been subdivided for statistical and local administrative purposes, through cooperation with a local committee in each case. Although this subdivision into tracts has been more or less arbitrary, several principles have been followed in laying out the tracts for each city. The tract areas are established with a view to approximate uniformity in population, with some consideration of uniformity in size, and with due regard for natural features. Each tract is designed to include an area fairly homogeneous in population characteristics. In cities where the ward lines are infrequently changed, the tracts may form subdivisions of the wards, but they are usually laid out without regard to the ward boundaries.

11. Block Statistics - Block statistics are tabulations of housing characteristics for areas as small as city blocks. They are available in separate reports for 209 cities which, in 1940, or in a subsequent census prior to 1950, had a population of 50,000 or more.

These publications are available at the Government Printing Office. Price lists are sent upon request to that agency.

12. Enumeration District - (ED) An enumeration district is the geographic unit of enumeration in the Census of Population and Housing. These districts are basically the territories laid out as work assignments for the census enumerators and are designed to provide statistics for each political or statistical type of area. In selected cities the area is further identified by block within the enumeration district. Enumeration districts are within political subdivisions.

ED maps may be obtained from the Geography Division, Bureau of the Census. The pertinent ED data for sampling use are map reproductions showing the boundaries of the ED's and photostatic copies of machine tabulations of population characteristics for each ED. Among these population characteristics is the number "All Heads, 14 yrs. +", which corresponds (though not exact) to the number of occupied dwelling units. Some of the maps used by the Census are copyrighted and cannot be reproduced. In these cases, maps must be supplied to the Geography Division.

Enumeration district boundary descriptions may be obtained at less cost than the ED maps. However, MCD and voting precinct lines often used in the descriptions do not appear on available maps.

C. AERIAL PHOTO SERVICE

1. Commodity Stabilizations Service

Western Laboratory
Performance and Aerial Photography Division
Commodity Stabilization Service
U. S. Department of Agriculture
167 West Second South, Salt Lake City 1, Utah

For photographs of these states:

| | | | |
|------------|------------|--------------|---------|
| Arizona | Kansas | North Dakota | Wyoming |
| California | Montana | Oregon | |
| Colorado | Nevada | Utah | |
| Idaho | New Mexico | Washington | |

Eastern Laboratory
Performance and Aerial Photography Division
Commodity Stabilization Service
U. S. Department of Agriculture
Washington 25, D. C.

For all other states.

2. Soil Conservation Service

Director, Cartographic Division
Soil Conservation Service
U. S. Department of Agriculture
Washington 25, D. C.