A TRAINING COURSE IN SAMPLING CONCEPTS FOR AGRICULTURAL SURVEYS

STATISTICAL REPORTING SERVICE

U.S. DEPARTMENT OF AGRICULTURE

SRS NO.21

A TRAINING COURSE IN

SAMPLING CONCEPTS FOR AGRICULTURAL SURVEYS

By

Harold F. Huddleston
Statistical Reporting Service
USDA

April 1976

(Revised November 1980)

PREFACE

This set of training materials on conducting agricultural statistical programs is the result of experience in training and consulting with officials of foreign governments, and reflects the needs encountered by personnel who have engaged in similar domestic assignments. The materials describe the sampling considerations commonly required in collecting current agricultural data and providing evaluations of agricultural census procedures. The approach used in presenting these materials is that of a discussion of the concepts followed by a relatively simple illustration. When possible, a real example is presented in which the participants complete a given unit of work. This approach is designed to satisfy the needs of survey personnel and administrators who will have responsibilities for carrying out surveys and who must be familiar with the essential concepts of sampling. It is believed that the demonstration of the interrelationships between different survey activities and alternative uses of information (that is, resources) should also be beneficial.

One of the main objectives in preparing this module was to provide a fairly complete set of materials for use in foreign training. A short training course consisting of 40 to 50 hours along with informal discussions over a period of two weeks is believed a feasible undertaking. The training materials are divided into four subsections as follows: (1) an overview of sampling, (2) construction of sampling frames, (3) random selection techniques, and (4) examples of agricultural surveys including the preparation of estimates.

It is necessary to emphasize that this training module is not intended to be a self-instruction manual, but rather an aid to be used by an instructor experienced in sampling principles. Also, it is assumed that the participants in charge of managing surveys will obtain additional training in sampling theory and survey design.

NOTE: (1997)
SCANNED VERSION ^ DOESN'T CONTAIN NUMEROUS COURSE MATERIALS SUCH AS MAPS, LISTINGS, PHOTOS, SATELLITE DATA, ETC. INCLUDED IN SCANNED DOCUMENTS FOR CONCEPTS ONLY - NO LONGER A TRAINING DOCUMENT.

ACKNOWLEDGEMENT

The author wishes to acknowledge the valuable contributions to this publication provided by several members of the Statistical Reporting Service's Research Division: Naomi Klaus, to area frame construction; Burgess Guinn, to the list materials; and Paul Hopkins, to the development of the computer programs. Encouragement to undertake the effort was provided by C. John Fliginger, formerly International Programs Officer for the Statistical Reporting Service.

CONTENTS

Page

CHAPTER 1 - AN OVERVIEW OF SAMPLING

1.1 Concepts Underlying the Sampling Procedure

Information is needed about a group or a universe of objects such as persons, farms, or firms. We examine only some of the objects and extend our findings to the whole group. There are four elements in the process: (1) constructing the frame to cover the population of interest, (2) selecting the sample, (3) collecting the information, and (4) making an inference about the population. These elements cannot generally be considered in isolation from one another. Frame construction, sample selection, data collection, and estimation are all interwoven and each has an impact on the others.

1.1.1 Information To Be Obtained

The information to be secured depends on the purpose for which the data are to be used. However, several basic concepts are required.

A. The universe to be sampled needs to be defined. Are we going to sample people, farms, households, etc.? In all cases, we shall be talking about a finite universe. That is, the group of objects or sampling units contained in the universe is limited.

B. For the universe defined, we shall be interested in one or more population characteristics which represent different sets of measurements to be obtained. These population characteristics correspond to content items on the questionnaire or reporting form. For example, for each of the universes mentioned above, we might be interested in the following population characteristics.

Universe of	Population Characteristics
People	Years of schooling; days with illness
Farms	Acres of corn; acres of wheat; number of cattle
Households	Number of persons in household; or income per household

C. Four common types of estimates are required:

(1) The population mean for the characteristic--such as persons per household.

(2) Population total for characteristic--such as acres of corn.

(3) Population proportion for characteristic--such as persons enrolled in school divided by total number of persons in population.

(4) Population ratio for two characteristics--such as income spent on food to value of housing unit, or quintals of maize harvested divided by hectares of maize harvested.

Each of these types of estimates is defined in terms of population quantities. Thus, we refer to them as population "parameters." We define each of these below in mathematical terms where the symbols y_i and x_i represent measurements of characteristics for an individual unit in a universe of N objects. In some cases the interest centers on whether the unit has a certain characteristic, in which case, $y_i(1)$ indicates the unit has the characteristic and $y_i(0)$ that it does not.

(1) Population Mean $\rightarrow \bar{Y} = \dfrac{\sum_{i=1}^{N} y_i}{N} = \dfrac{Y}{N}$, Hectares of Maize per Sampling Unit.

(2) Population Total $\rightarrow Y = N\bar{Y}$, Hectares of Maize for All Sampling Units.

(3) Population Proportion $\rightarrow P = \dfrac{\sum_{i=1}^{N} y_i(1)}{N} = \dfrac{N_S}{N}$, $\dfrac{\text{Number of persons enrolled in school}}{\text{Total number of persons}}$.

$$\sum_{i=1}^{N} y_i(1 \text{ or } 0)$$

(4) Population Ratio $\rightarrow R = \dfrac{\sum_{i=1}^{N} y_i}{N} = \dfrac{Y}{X} = \dfrac{N\bar{Y}}{N\bar{X}} = \dfrac{\bar{Y}}{\bar{X}}$, $\dfrac{\text{Quintals of Maize Harvested}}{\text{Hectares of Maize Harvested}}$.

$$\sum_{i=1}^{N} x_i$$

The purpose of sampling is to provide estimates of these parameters based on a sample of the units from the universe. The estimates obtained from the sample are referred to as "statistics." The mathematical form used to provide the estimate from the sample is referred to as the "estimator." Many of the estimators look very much like the population parameter, i.e., they are "copies" of the parameter. However, modern sampling theory has developed many alternative estimators whose properties need to be known to avoid their indiscriminate use.

1.1.2 How Good Is Our Information?

A. Sampling Errors

For whatever type of estimate we may be interested in, we hope the "sampling error" will be small. This we measure primarily by the concentration of the sample estimates around their expected value for a hypothetical population. The expected value is the mean value of all the possible estimates based on a given estimator and sample size. This measure of concentration is provided by the sampling error of the estimator. Actually, it is not necessary to draw all possible samples to get a measure of the extent by which sample estimates differ from the expected value. By using sampling theory, it can be shown that, in simple random samples of size n (fixed sample size), the population variance of the sample mean $\bar{y}$ for selection without replacement is given by

$$V(\bar{y}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S_y^2 \text{ , where } \left(1 - \frac{n}{N}\right) \text{ is the finite population}$$

correction factor,

$$S_y^2 = \frac{1}{N-1}\sum^{N}(y_i - \bar{Y})^2 \text{ , and } \bar{Y} = \frac{1}{N}\sum^{N} y_i \text{ .}$$

The sampling error of the mean is frequently needed, and is defined as the positive square root of the variance of the estimate. In addition, the relative sampling error is commonly used and referred to as the coefficient of variation of $\bar{y}$, i.e. $CV(\bar{y})$. Thus, there are several ways of expressing the precision of the estimate.

(1) The relative error is:

$$CV(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{\bar{y}} \text{ .}$$

(2) Another way of expressing the precision of our estimate is by use of confidence intervals for normally distributed means, such as:

$$\bar{y} \pm 2\sqrt{V(\bar{y})} \text{ ,}$$

$$\text{or} \quad \bar{y} \pm 3\sqrt{V(\bar{y})} \text{ .}$$

The choice of the multiplier 2 or 3 depends on the degree of certainty with which we desire to make statements about the interval containing the

population mean. The probabilities associated with the confidence intervals are called confidence coefficients. In the discussions in Section 1.2, we will use the range of the estimates rather than the sampling error for comparative purposes. The range is an efficient measure of variability for small samples ($n \leq 20$), consequently it will be satisfactory for our purpose and easier to obtain.

B. Nonsampling Errors

It is too commonly assumed that there are no reporting errors in the data collected. Actually, errors of measurement or responses are almost always present when information is collected. The problem is how to form useful estimates from the sample in the presence of such errors. Let us start with a simple model to make the ideas clear. If the true figure is y, we will observe or measure a quantity $\dot{y} = y + e$ where e is the nonsampling error with $E(e) = 0$ and $V(e) = K$. When we take a sample of establishments, the expected value of the sample mean will be $E(\bar{y}) + E(\bar{e}) = \bar{Y}$ (the population mean). But the error of the estimate is: $V(\bar{y}) + V(\bar{e})$ which is larger than the variance obtained when no measurements or response errors are present. In this case, the sample mean is unbiased but its error has increased. If, however, the errors are deliberate on the part of the respondent, $E(e)$ will not be 0 and the estimate is subject to a bias. Frequently, larger establishments will understate the value of the characteristic we are interested in and the smaller ones will also understate the true value. In addition, the data collected from the same establishment will differ from interviewer to interviewer, since interviewers have their own concepts and whims about surveys. In this case, we have a very difficult situation confronting us. The influence of the interviewers is commonly controlled by using replicated samples and assigning each interviewer one or more sampling units in each sample. Replicated sampling is discussed in more detail later. Further, an attempt is made to minimize these errors by questionnaire design and enumerator training. Thus, we can see there are many kinds of errors involved when data are collected from a sample of objects or from the whole group.

Suppose we want to estimate the average number of employees of a universe of business firms by taking a sample of business firms. Some firms may not know the exact number of employees working on a given date;

some may overstate it as a matter of prestige; a few may refuse to give the information; in a few cases, the enumerator may record the answer incorrectly; in processing the data errors will be made; etc. Thus, errors of various types may creep into the results. These errors are present whether you take a sample or interview every unit in the universe (census). There are also errors arising from unclear definitions either in the questionnaire or in the instructions to the enumerators. Also, you may not define all the terms with mathematical exactitude and so you may not know exactly what you want to measure. This brings about errors.

Some of the errors in the data are of the random type; that is, they average to zero over the sample. This is generally the case when the errors are not deliberate or intentional. Some units will be overstated and others will be understated, resulting in a net difference near 0. There are other errors of a systematic type which are more serious. This type of error will not cancel out over the sample, but will persist. Such errors are called systematic errors.

C. Sample Size

It is generally assumed that the larger the sample the more reliable our information will be. Certainly this is true for errors due to sampling and technical biases in the estimator used. However, for most sources of nonsampling errors the reverse is true: the larger the sample becomes, the more likely additional errors are to be present and to be serious.

Sampling technicians are generally fully aware of sampling errors and biases due to the estimator used, but may be unfamiliar with or unable to assess the impact of nonsampling errors on the survey. Also, the data user may be unable to evaluate the seriousness of nonsampling errors. Thus, it becomes almost mandatory to either pretest the survey procedures the first time it is conducted or to do a postsurvey evaluation to measure the extent of these errors. Likewise, surveys which are repeated at given intervals of time may be expected to be subject to additional sources of nonsampling errors due to longevity. Thus, it is essential to provide resources to keep these various sources of errors in balance. In the typical sample survey, sampling errors should be the major sources of error, while in the typical census undertaking nonsampling errors are almost always the major source of errors. Thus, the choice between a

sample and a complete census is primarily based on considerations other than the level of errors, such as data needs and objectives:

(1) How soon are the data needed?

(2) How much detail is required for subuniverses or regions?

(3) How much money is available?

(4) How large and what type of organization is available to manage the work?

The control of sampling errors requires previous knowledge of parameters such as means, variances, ranges, or shape of distribution for population characteristics.

We shall not examine the use of all these sources for determining sample size at this time, but merely point out one way of using knowledge of the coefficient of variation. How many sample units n should be taken from a universe of N units? If previous information indicates what $\bar{Y}$ and S_Y are, then the CV(y) can be used to estimate the approximate sample size required, since

$$CV(y)\ \left[\frac{1}{n}\left(1 - \frac{n}{N}\right)\right]^{\frac{1}{2}} = \text{Relative Error Allowable}$$

where the CV(y) and the relative error are stated in the same terms, i.e., either as percentages or as decimal fractions.

The purpose of sample survey design is to find ways of using information and sampling theory to reduce the sample size or the errors in the estimate through the selection technique, estimation method, or grouping the universe units into strata.

1.1.3 Use of Information in the Application of Sampling Theory

Collecting data is dependent upon using sampling theory. However, the application of sampling theory requires some knowledge of means, variances, universe size, and other parameters if an efficient survey plan is to be employed. If such knowledge is not available from earlier surveys or censuses, the plan will almost surely not be as efficient as could be devised with it. However, valuable information may exist in the form of what is referred to as "supplementary information," or "auxiliary information." It is not unusual to find that some information is already available for the various units composing the universe. For example: (1) the number of inhabitants (or farms) in different villages may be known from a previous census, or (2) geographical areas

based on cadastral maps may be determined, or (3) areas by broad categories of land use may be available from photographs. It is important to make use of this information for improving the precision of the estimates. This supplementary information about some variable x may be used in a variety of ways, such as:

A. Selection of sampling units with probability proportional to x if information is available for the individual units, or

B. The available information may be used to form different estimators, such as:

(1) Reciprocal of probability of selection estimate (requires N)

(2) Ratio type estimate,
(3) Difference estimate,
(4) Regression estimate,
} (require auxiliary variable(s))

C. Stratification of units into groups on the basis of information about x. An attempt is made to make the strata (or groups) internally homogeneous by placing in the same stratum units which appear to be similar. For this purpose it is not necessary that quantitative values of x be available, but only that the units be similar.

1.2 Survey Strategies Based on Auxiliary Information

1.2.1 Method of Selection

A. Equal Probability Sampling

(1) The selection of each unit with the same probability constitutes the fundamental method of sample selection. From a universe of N units select one by giving equal probability $(\frac{1}{N})$ to all units. Make a record of the unit selected and return it to the universe. If this operation is performed n times, we get a simple random sample of n units, selected with replacement.

(2) If this procedure is continued till n distinct units are selected but all repetitions are ignored, a simple random sample of n units, selected without replacement is obtained. An alternative procedure is to select the first unit with probability $\frac{1}{N}$, the second unit with probability $\frac{1}{N-1}$, the third unit with probability $\frac{1}{N-2}$, ..., the last unit with probability $\frac{1}{N-n+1}$. It can be shown that for either of these schemes of nonreplacement sampling that the probability of a specified sample of n units, ignoring order, is $\frac{1}{\binom{N}{n}}$.

The estimation of the sample mean for with replacement and non-replacement sampling is the same, but the sampling error differs by the constant multiplier known as the "finite population correction factor." This factor is $(1 - \frac{1}{N})$ for with replacement sampling and $(1 - \frac{n}{N})$ for nonreplacement sampling. The units are selected for both schemes by use of a random-number table containing the integers from 1 to N.

B. Unequal Probability of Sampling

Another method which can often be used to achieve greater concentration of the sample estimates around the expected value is to make use of information available for a variable x for each of the N units. It is necessary that the variable x be positively correlated with the characteristic being estimated. The following example of a universe of units is given in table 1 below. The units are selected by use of a table of random numbers containing the integers from 1 to X = 250. Thus, the probability of selecting unit #3 is greater than that of the other units.

The reciprocal of the probability estimator of the mean based on a single unit is: $y_i = P_i N$ for either EP or UEP sampling. It is clear from inspecting the last column on the right that unequal probability selection, in this case, leads to a greater concentration of the estimates around the mean based on samples of size 1. The range of the estimates for EP sampling is 9-67 and for UEP 22.50-30.00. The expected value for both methods is 27.0 and has the property of being unbiased ever though the individual unit estimates are averaged using different sets of probabilities.

Table 1--Small universe of 10 firms giving number of employees per firm

Unit label L	Auxiliary information	Probability of selecting unit on first draw		Number employees on survey date	Estimate of population mean from unit	
	x = No. employees 5 yrs. ago	Equal probability	Unequal probability based on X	y_i	EP	UEP
1	30	$\frac{1}{N} = .1$	$\frac{30}{250} = .120$	31	31	25.83
2	15	$\frac{1}{N} = .1$	$\frac{15}{250} = .060$	15	15	25.00
3	60	$\frac{1}{N} = .1$	$\frac{60}{250} = .240$	67	67	27.92
4	18	$\frac{1}{N} = .1$	$\frac{18}{250} = .072$	20	20	27.78
5	12	$\frac{1}{N} = .1$	$\frac{12}{250} = .048$	13	13	27.08
6	15	$\frac{1}{N} = .1$	$\frac{15}{250} = .060$	18	18	30.00
7	10	$\frac{1}{N} = .1$	$\frac{10}{250} = .040$	9	9	22.50
8	20	$\frac{1}{N} = .1$	$\frac{20}{250} = .080$	22	22	27.50
9	45	$\frac{1}{N} = .1$	$\frac{45}{250} = .180$	48	48	26.67
10	25	$\frac{1}{N} = .1$	$\frac{25}{250} = .100$	27	27	27.00
Total N = 10	X = 250	1.0	1.0	Y = 270		

1.2.2 Method of Estimation Using Auxiliary Information

A. Difference Estimation

The auxiliary information on x is used to estimate the change $\bar{y} - \bar{x}$ and this is added to $\bar{X}$, the population mean for the auxiliary variable. Since x and y need not be the same characteristic measured on two different dates, the

general form for the difference estimator is as follows:

$\hat{\bar{Y}} = \bar{y} - K(\bar{x} - \bar{X})$; or for K = 1, when x and y represent the same characteristic, $\hat{\bar{Y}} = \bar{X} + (\bar{y} - \bar{x})$

where any value of K may be used (which is determined independently of the survey data) but in our example K = 1 since x and y represent the same characteristic. For all estimates obtained from table 1 using the EP method of selection, $E(\bar{Y}) = 27.0$, which indicates the difference estimator is unbiased, with the range of the estimates being 24-32.

B. Ratio Estimation

Instead of estimating the difference between $\bar{y}$ and $\bar{x}$ we may estimate the ratio of the means from the sample and multiply it with the known value of $\bar{X}$. The estimator

$\hat{\bar{Y}} = \bar{X}\frac{\bar{y}}{\bar{x}}$ is called the "ratio estimator," assuming that $\bar{x} \neq 0$.

The estimates are the same as those given in the right-hand column of table 1 when the EP selection method is used. Since all units were selected with EP, the average of all estimates is 26.728, which shows that this estimator is not unbiased. However, the range of the estimates is 22.50-30.00 which is the same as for the sample mean using the selection method of UEP.

C. Regression Estimation

Instead of making use of the value K (which is independent of the survey) in the difference estimator, we calculate the regression coefficient b from the sample and use this in place of K. The estimator is $\hat{\bar{Y}} = \bar{y} - b(\bar{x} - \bar{X})$. However, we cannot use this estimator for a sample of size n = 1, but require $n \geq 3$. This estimator is expected to be more precise than the difference or the ratio estimators, but it is more cumbersome to calculate. In addition, only approximate formulas for its bias and variance are known. If "b" were known or available for a large sample, the estimates for individual units could be computed. For this small universe, "b" = .9976, and the estimates derived using this value give a range of 23.96-32.08, based on n = 1.

1.2.3 Stratification of Units

Stratification implies that the units in the universe are grouped (or stratified) on the basis of information about some quantitative or qualitative variable x. An attempt is made to make the strata internally homogeneous by placing in the same stratum units which appear to be similar. Then by selecting a sample of a suitable size from each stratum, it is possible to produce an

estimate which has considerably smaller sampling error than that given by a simple random sample from the entire population. The 10 units (table 1) are divided into two strata of 5 units each, based on the X values. The smallest units: 2, 4, 5, 6, 7 are placed in the first stratum and the remaining units in the second stratum. The range of the estimates based on a simple random sample of one unit from each stratum is 15.5-43.5. If a simple random sample of size two is taken from the universe of 10 units, the range is 11-57.5. The results of 1.2.1, 1.2.2, and 1.2.3 are summarized in table 2.

1.2.4 Exercises 1 and 2

Exercise 1:

Each student is to draw all possible random samples of size 2 from the population of 10 units.

(a) Using EP sampling and the difference and ratio estimators, determine the range of all possible estimates of the population total for these two estimators.

Exercise 2:

Each student is to draw all possible random samples of sizes 1 and 2 from the population stratified into two strata.

(a) Same as a) above.

(b) Using UEP sampling and the reciprocal of the probability estimator, determine the range of all possible estimates of the population total, ignoring the order of selecting the two units (i.e., there are 20 possible samples, 5 x 4, in each stratum).

Table 2--Comparison of range of estimates for different sampling strategies

Type of strategy used	Expected value of mean	Range of estimates	
		n = 1	n = 2
Sampling from entire universe:			
EP - Estimator 1	27.0	9 - 67	11.0 - 57.5
UEP - Estimator 1	27.0	22.5 - 30.0	24.62 - 29.36
EP - Estimator 2	27.0	24 - 32	1/
EP - Estimator 3	26.728	22.5 - 30.0	1/
EP - Estimator 4	3/ 27.0	2/23.96 - 32.08	2/24.966 - 30.066
Sampling from two strata:			
EP - Estimator 1	27.0	---	15.5 - 43.5
UEP - Estimator 1	1/		1/
EP - Estimator 2	1/		1/
EP - Estimator 3	1/		1/
EP - Estimator 4	2/		2/

1/ To be completed by students before next session.

2/ The sample is too small to compute the slope without an independent estimate of "b".

3/ If the estimates were calculated based entirely on sample data, the estimator would be slightly biased.

1.3 Acquisition of the Sampling Frame

A sampling frame is a means of gaining access to the universe we are interested in sampling for factual information about one or more population characteristics. The sampling frame is composed of units (frame units) which may or may not be the same as the units in the universe. Consequently, we have a basic problem of developing a linkage or building a bridge between these different types of "units." We call this linkage the "survey operational rules," which are required for a particular survey based upon a particular sampling frame.

Clearly, the most important step in sampling is acquiring the frame, since without it we cannot apply the principles of survey design.

1.3.1 Characteristics of Frames

An ideal sampling frame is a list of distinct, clearly defined, mutually exclusive sampling units containing all the elements of a specified universe.

The individual sampling units may be natural units, artificially constructed units, or some convenient reporting or working unit. It is not always necessary to have a complete listing of individual sampling units. Clusters of units may be used provided cluster sizes (the number of individual sampling units contained in each cluster) are known and procedures developed for an unambiguous definition of the individual sampling units within clusters.

We may characterize sampling frames in terms of their defects. It is important that we keep these in mind when constructing and using frames. Characterization of frames include the following: (a) A frame is termed inaccurate if the units listed are incorrectly or imprecisely defined or if information pertaining to the units is inaccurate; (b) a frame is incomplete if any units of the population are omitted; (c) a frame contains duplication if some units are included more than once; (d) a frame is inadequate when it does not cover all the universe of interest in a particular survey; and (e) a frame is out of date when it no longer reflects the universe, although it may have been accurate, complete, and free from duplication at the time of construction.

(a) Inaccuracies in frame definitions of sampling units should be discovered during the course of a well-designed survey, and sample data may be adjusted so that valid inferences result. If control information is inaccurate, the efficiency of a sample will be reduced, but bias is not necessarily introduced.

(b) Incompleteness in the frame results in the exclusion or omission of part of the universe. It will usually not be discovered during a survey. Incompleteness is often more serious than it appears at first sight, since it is often confined to units possessing some special characteristic, which may be seriously underrepresented in the sample.

(c) Duplication has an opposite effect from incompleteness since the duplicated units have more than one chance of being drawn. However, this will almost always be a tedious operation, since it requires a very careful review of each individual sampling unit.

(d) Inadequacy in the frame will usually be known before surveys are undertaken from the specifications of the frame itself. Inadequacy can and must be dealt with by construction of subsidiary frames for omitted categories or groups.

(e) Out-of-date frames are likely to be found in sampling situations where the basic sampling units may appear or disappear from the universe. The resulting deterioration is extremely troublesome in list frames but is not an important factor for area frames. The only way to minimize this defect is to review the frame and bring it up to date periodically.

1.3.2 Source Materials for Construction of Sampling Frames

Auxiliary information may sometimes be acquired in the frame construction. Frequently the source materials contain valuable information which can be used. Examples of source materials for frame construction

(a) A map of land area based on county maps;

(b) Photographs of land area;

(c) City directory;

(d) Lists of blocks in a city or subdivision;

(e) List of members of a trade association;

(f) List of farms from last census;

(g) List of participants in government programs;

(h) Telephone directory.

1.3.3 Types of Frames Used in Sampling Work

A. Population and Housing Census

Such frames are based on listing places of abode and have to be brought up to date in order to take in the new construction of the conversion of building for use as housing. The usefulness of such frames is enhanced if a sample of areas is selected from it at the time the census is taken, or the census is tabulated by areas.

B. Town Plans

Street maps of individual towns may provide a suitable access to certain universes of interest.

C. Lists of Villages

In countries where households are clustered in villages, a listing of villages, may provide a satisfactory frame of households.

D. Directories of Establishments

Frequently, a business association, census or licensing of firms can be used as a frame for certain kinds of businesses.

E. Area Frames

Generally, area frames are the most permanent frames, since the land area changes very slowly in all countries. Two types of source materials are found useful.

(1) Maps showing parcels of land by ownership.

(2) Special purpose maps showing roads, elevation, soil types, vegetative types, etc. In addition, photographs of a country provide another important source.

2.1 Introduction

We have discussed some general concepts useful in sampling and the necessity of a sampling frame in carrying out a survey or census. We now turn to the task of constructing sampling frames.

We will consider construction of two types of frames--area and list. The sampling unit in an area frame is some specific area of land; the sampling unit in a list frame is usually represented by a name or address, or both. The major advantage of an area frame for agricultural purposes is that it is complete, the entire universe being contained in the frame. A list frame, on the other hand, is to some degree incomplete by the time it is assembled, since farms and firms are continually being formed and dissolved. One of the greatest difficulties with either type of frame is clearly and unambiguously defining the sampling units or elements in the frame. Success is measured in terms of the following characteristics:

1. Accuracy - In an area frame, a map with poorly drawn features may result in an inaccurate measurement of a sampling unit, or an unequal division of sampling units.

2. Freedom from duplication - If units are indefinitely delineated in an area frame, the same area might be included in more than one sampling unit. Duplication in a list frame may result from one element being listed twice with different descriptions (e.g., an agricultural operation listed by the name of the owner and the name of the farm).

3. Completeness - A list frame is almost never complete because membership of a group is continually in flux due to additions and departures. One of the advantages of an area frame is that it is complete and the completeness can be verified by inspection of mapping materials available in a central location.

4. Timeliness - Area frames generally can be constructed with minimal or no field work and remain up to date for a relatively long period because of the difficulty and expense involved in changing terrain. In contrast, the elements in a list frame are hard to keep up to date.

It is of primary importance in constructing sampling frames to define sampling units which both the interviewer and the respondent can identify. There are many

available sources of information which can be used for frames -- soil maps, topographic maps, aerial photography, maps of population density, census enumeration or supervisor's work areas, etc., for area frames; and census listings, telephone directories, brand registrations, membership lists for industrial or commodity organizations, and so forth, for list frames.

Sampling error can generally be reduced if the frame is stratified. The object of stratification is to place sampling units into groups which are as alike as possible within groups and as different as possible between groups. A list frame might be stratified on the basis of income, size of farm, address, etc., depending on the information available. In an area frame, population density is a common criterion. Or, in a frame used to make agricultural estimates, stratification could be based on crop or land use. It should be noted that while land-use stratification is useful for broad categories, it is a highly subjective process which does not lend itself to detailed, specific classification. Also, because frames are expensive to construct, they should be general enough to accommodate different types of surveys for a number of years even though the immediate needs may be for specialized surveys.

2.2 Construction of an Area Frame for the Dominican Republic

As an exercise you will be asked to construct an area frame stratified according to land use for the Province of Salcedo in the Dominican Republic. You will subsequently select a sample using this frame. The frame will be constructed in two major steps: first, classification of the land according to its use; and second, construction of sampling units. Frequently, it is found in constructing sampling units that the best use of resources (time and personnel) will indicate that only very large primary frame units should be defined and that these should be classified into strata. Then, only primary units selected in a sample are subdivided into elementary units. Thus, two types of units are defined for the frame: (1) Primary frame units (i.e., "count units") and (2) elementary frame units (i.e., sampling units). You will be using topographic maps as your primary resource in the construction, but the following other materials are also available: aerial photographs, maps of soil type, population density, and vegetation; and agricultural census data by municipalities.

2.2.1 Stratification Based on Land Use

Five land use groups were defined for the Dominican Republic and are given later in this section. All five strata are present in Salcedo Province. Because

coffee is an important crop and the first survey was to be a coffee survey, land where coffee (and also cacao) was known to be planted was put in a separate stratum. Generally, it is not possible or desirable to stratify in such detail for specific crops, but coffee and cacao are fairly permanent crops and unlikely to change rapidly with time. The following materials are included in training set #1 with which you will construct a sampling frame for one province.

1. Two blank topographic maps covering the Province of Salcedo.
2. A sketch showing the strata constructed in 1972.
3. Transparent overlays showing count units constructed in 1972.
4. Two blank sheets for listing count units.

The materials in items 2 and 3 are to be used by you to compare your work when finished with that done in 1972. There is no single or unique solution to the construction of a sampling frame; it is a matter of using the best collective judgment available.

In delineating strata, use of natural physical boundaries which can be easily identified on the ground is a primary consideration. Even though this may mean including some land which may not conform exactly to the land use of stratum definition, this is preferable to a boundary which cannot be identified correctly by the interviewer. Check the legend on the maps so you will recognize the different kinds of culture, roads, dwellings, land types indicated by the colors, etc.

Do not isolate very small areas of land which seem to conform to a different land use definition. As a general rule, no less than four square kilometers (two sampling units) should be separated.

Large bodies of water may be removed from the frame and may need to be considered as a special stratum if houseboats or fish farms are commonly encountered. Any bodies of water that are greater than one square kilometer should be outlined in the color of the stratum surrounding them and labeled "out" or "special stratum."

The following steps should then be performed:

A. Outline the province boundary in black.

B. Outline the cities. These areas should be large enough to show some street pattern, not just a crossroad with a few houses. Draw the boundary in green and shade it with yellow ochre.

C. The province which you are stratifying is an important coffee-producing area. Locate the areas designated "cafe" or "cafe and cacao." Delineate these with a blue line and shade the inside edge of the boundary.

D. Delineate the intensively cultivated areas in purple. These should include most places where a crop other than coffee is indicated. On photographs, these are the areas that would be expected to show a large proportion of fields. On the topographic maps, these are the areas that have some specific crop indicated or the color code indicates it is cropland. More than half the area classified in this stratum should be labeled with some specific crop on the maps (see the legend on the map).

E. Consider the remaining areas on the map and distinguish between III and IV. The best indications will be the amount of cleared land and the number of houses. Stratum IV should have a negligible number of houses, very little cleared land. It is frequently mountainous areas. Stratum III should be drawn and shaded in green, Stratum IV in orange.

F. After you have finished stratifying, put the maps together and be sure that:
   a. the stratum boundaries are continuous from page to page;
   b. all areas have been stratified; and
   c. all outlined areas meet the stratum definitions.

An alternative training set (training set #3) is also included for a situation where aerial photographs are available for stratification. These materials provide a second example which can be completed after training set #1 is finished.

Land Use Definitions

Stratum I purple	Intensive Agriculture This stratum is cultivated land. Many of the crops will be irrigated. Include tree crops, exclusive of coffee, such as bananas, cacao, and plantain. Other crops will include rice, sugarcane, maize, tobacco, beans, and peanuts.
Stratum II blue	Coffee and Cacao This stratum is land in tree cover, devoted primarily to the production of coffee and cacao above 100 meters. Limit this as much as possible to land where coffee is specifically indicated. This stratum may include some cacao where it is interplanted.
Stratum III green	Extensive Agriculture This is a mixture of cropland and cleared land used for grazing livestock. This is the most loosely defined stratum and, in practice, it is land which fits none of the other definitions.
Stratum IV orange	Nonagricultural Land This is land in natural cover which supports very little or no agricultural activity. This will include mountains, forest and swamps.
Stratum V yellow ochre	Urban This is concentrations of population ranging from small rural towns to major cities.

2.2.2 Construction of Primary Frame Units

To identify all the sampling units in an area frame would be very time-consuming and therefore expensive. If, however, the number of sampling units in the frame is known and the specific area in which they are located is known, then we can avoid splitting the area into individual sampling units. One way of accomplishing this is by constructing primary frame units. A primary frame unit is a specific area of land containing an assigned number of elementary frame units (i.e., sampling units). Each primary unit has a given number of sampling units. The primary frame units are commonly called "count" units, because a count of the number of sampling units is available. The area in each count unit is measured and this area is divided by the expected sampling-unit size to derive the number of sampling units. A tolerance is allowed around the expected size to permit the use of physical boundaries and allow the number of sampling units assigned to a count unit to be rounded to a whole number.

When the universe has been completely divided into count units, the count units measured, and sampling units assigned, a list is made of all the count units, identifying them by location and number of sampling units. This list is the frame from which samples are taken. When a sample is selected, only those count units containing selected sampling units are split into individual sampling units. However, it should be pointed out that, if only the total number of count units in the frame are known rather than the total number of sampling units in the frame, unbiased estimates can be made by using multistage sampling. This is frequently the case when sampling units are identified in the field rather than using mapping materials, especially maps or photographs which are quite old.

Construct count units for each stratum in Salcedo Province except Stratum V.

A. Boundaries

   Count-unit boundaries should be physical, permanent boundaries which can be identified on the ground, such as roads, rivers, railroads, etc. Draw count-unit boundaries in the color of the stratum, but do not shade them.

B. Size

   Try to keep the count units within the desired size range. However, this is not as important as using a good boundary. The size restriction on count units is to make them easier to split at the time of

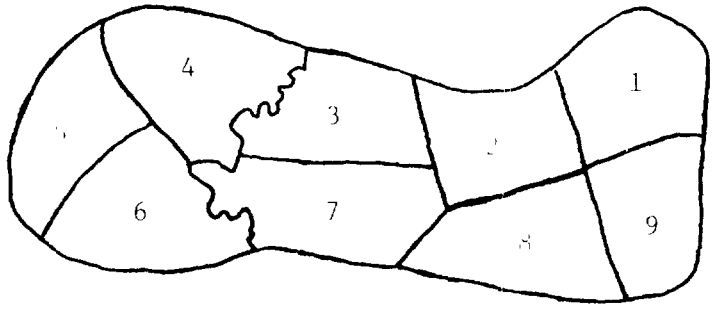
sample selection, but they should be large enough to accommodate a larger sampling-unit size should this be desirable for subsequent surveys.

Stratum	S.U. Size	C.U. Size
I, II	2 km^2	12 - 24 km^2
III, IV	4 km^2	24 - 48 km^2

C. Identification

Each count unit is identified by the number of the stratum in which it is located, a count-unit number, and its area. The area is used to compute the number of sampling units to assign to the count unit.

1. Number the count units in a serpentine fashion beginning in the northeast corner and proceeding from east to west. This is done in an effort to standardize and eliminate any bias which might be introduced in numbering the count units. However, the consecutive numbering of count units with similar agriculture is commonly employed in constructing "paper" strata. (See Section 4.2.4 for discussion.)



   a. Count units are numbered within each stratum. There is a count unit numbered 1 for each stratum.

   b. Put all the map pages together and number the count units throughout the province.

2. Using a grid (the one on the maps or a plastic overlay), estimate the number of square kilometers in each count unit. In actual practice, a planimeter is used to obtain a more accurate measure of the area. Salcedo contains approximately 420 square kilometers.

3. Write the identification of each count unit on the map in the color of the stratum.
   For example: I - 3 - 16
   This is count unit 3 in Stratum I and it contains 16 square kilometers.

4. List the count units on the count-unit identification sheets, preferably a separate sheet for each stratum. Include the map page number, the stratum number, the count-unit number, and the measurement on these forms.

2.2.3 Assembling Frame Units for Sampling

After the count-unit identification sheets have been made, the count units can be assembled in whichever way best suits the needs of a particular survey. In the Dominican Republic, there was interest in providing information for five regions in the country, so count units were grouped by stratum and province in geographic regions. Grouping could also be based on types of crops grown, size of farms in the area, number of crops in a season, etc.

2.3 Use of Satellite Photography in Constructing an Area Frame

The launching of the first Earth Resources Technology Satellite (ERTS) in July 1972 has provided an additional source of information on land use. Each ERTS (now called LANDSAT) image (or "frame") covers an area of 100 nautical miles by 100 nautical miles. The scale of the 9" x 9" photographs is approximately 1:1,000,000. Monthly U.S. and non-U.S. Standard Catalogs and microfilm copies of the pictures are available showing the areas for which imagery is available along with certain additional information, of which the percentage of cloud cover is most important. These catalogs can be purchased from the U.S. Government Printing Office, Washington, D. C. 20402, and the microfilm copies may be ordered through the EROS Data Center, Sioux Falls, South Dakota 47198.

An enlargement for an area covering the Province of Monte Cristi in the Dominican Republic is included in training set #2. The enlargement is at a scale of approximately 1:250,000. We shall use this enlargement and transfer from the topographic map the more important roads, rivers, and other natural features shown on the map for reference purposes. Having transferred the main features identified on the map, we shall proceed to construct the strata for an area sampling frame using the enlarged ERTS image. In training set #2 are the transparent overlays showing how the stratification and frame was constructed in 1972, without the ERTS picture on one map page.

2.3.1 Stratification of Primary Land Use Features

The stratification of the land area into broad land use categories which may be helpful for agriculture is based on what the photo interpreter can see. Obviously, the photo interpreter's training and experience are important factors in how well important strata can be delineated. Also, it is quite important that the land use be known for a number of selected areas within the ERTS picture in order to do the best job of stratification. For the material enclosed this additional information was not used. An example of the stratification is shown in training set #2. The materials included correspond to:

1. A portion of an ERTS photo covering Monte Cristi Province.
   28AUG72 C N20-05-W070-48 N N20-03/W070-41 MSS 5
2. Transparent overlay showing one possible land-use stratification.
3. A topographic map covering the Province of Monte Cristi.
4. Transparent overlays showing strata constructed in 1972.

2.4 List Frame Construction

An ideal list frame would consist of all the attributes for a frame described earlier. That is, the list frame should consist of distinct, clearly defined, mutually exclusive sampling units containing 100 percent of the elements in the population that it is desired to have an estimate for. Such an ideal frame rarely exists for sampling or census work relating to agriculture. Normally, the list frames which must be used for agricultural purposes are out of date and incomplete. Generally, such lists were not assembled for agricultural sampling but were the by-product of a program with a different objective. Consequently, the list constitutes an imperfect frame.

An example of a list frame follows which is designed to give the reader some insight into problems associated with a typical list frame in the United States. The fact that very few "natural" list frames for agricultural sampling exist necessitates a certain amount of ingenuity in developing a frame that is operationally useful as well as valid and efficient for sampling. Incomplete or specialized lists are commonly combined to form relatively complete frames which can be used with an area frame employing multiple-frame theory.

2.4.1 List Frame Development

There are several phases of developing and maintaining an efficient list sampling frame that contains a large portion of the population of interest.

The first and basic phase is identifying the best list source if there are many sources available. This may be extremely critical because the task of merging several lists together is difficult. This "best" list source should provide a list of farm and ranch operators which represents a high proportion of each province's crops or livestock.

The list source should also provide control information for purposes of stratification. A large list of farm operators without control information is no more efficient than an area sampling frame which is not stratified. The most important piece of information about each list unit is whether or not the "name" has the survey item(s) of interest. It is also important to know the relative size of each unit with the crop or livestock species. This size can consist of: (1) actual inventory numbers, peak numbers during the year, or marketings during a period of time; (2) an index created from more than one data item, or; (3) a size code indicating relative size of operation.

The list frame must provide means to select samples with known probabilities. Names on the list have to be associated with an identifiable operating unit. Also, duplicated names must be removed when developing the list. If several list sources are to be used, consideration must be given to increasing problems of duplication removal. The task of identifying duplication in a composite list is not an easy one.

Defining operational arrangements (i.e., people and business arrangements) is quite desirable to the extent possible. This is helpful in identifying duplication and in the classification of parcels of land that could be part of several operations. Information on corporate or ranch names and names of all associated partners is invaluable in applying appropriate procedures necessary for unbiased multiple frame estimates.

2.4.2 Steps in Frame Construction Based on Name-Address Lists

A. Name and address units are assembled from the available list sources to make the frame as complete and up to date as possible.

B. Identify and remove duplication of units with the same name and address.

C. Determine procedures for associating frame sampling units (i.e., unique names and address) and the population units to be surveyed for information, i.e., farms, households, etc.

D. Arrange list for sampling by strata and assign each unique frame unit (each line in the list unit identification sheet) a number to be used in random selection.

2.4.3 Exercises 3 and 4

At the end of this chapter there is an alphabetic computer listing (retyped) of names, addresses, auxiliary information and showing duplicated units removed by crossing out the name-address unit. The original list of 397 name-address units was reduced to 269 name-address units. The column headings indicate the information shown for each name-address unit.

Exercise 3:

Each student is to structure the list so the frame consists of clusters based on the city within state (columns 6 and 7) or, if no city is provided, the county within state (column 3) in place of the city. After the clusters are formed, record the number of name-address units in each cluster.

Exercise 4:

Each student is to create an auxiliary variable(s) for each name-address unit for use in estimating livestock based on the livestock auxiliary data in columns 8, 9, 10. After you have created the variable(s), you should stratify the name-address units into no more than 12 strata. Since the auxiliary data could also be used for unequal probabilities of selection, or for a ratio estimator, it is usually desirable to require the auxiliary variable(s) x_i be greater than zero. If there is no data in columns 8, 9, or 10 (i.e., a zero), x_i can be set equal to 1. If this is done, then the ratio is defined for each unit and division by zero will be avoided no matter what size sample is selected later.

Two principle criteria are to be used in constructing the strata after ordering the units from smallest to largest: either (1) the total measure of size for all name-address units should be approximately the same for each stratum, or (2) the square root of the largest measure of size should be divided by the number of strata desired to yield the width of the strata: i.e., $W = \sqrt{\text{largest } x_i} : S$. Name-address units are then placed in strata based on the magnitude of the square root of x_i.

CHAPTER 3 - RANDOM SELECTION TECHNIQUES

3.1 Introduction

After the sampling frame has been constructed, we need to develop techniques for selecting samples so we may make inference about the totality of all units in our population of interest. In chapter 2 it was indicated that the count units (or elementary frame units for the list frame) could be assembled in whichever way best suits the needs of a particular data collection program. We then developed a labeling of the sampling units in the frame so they could be identified by the integers from 1 to N. However, any arbitrariness for the sake of convenience, or special purpose in listing of the frame units, may introduce unpredictable biases into the ordering scheme. This can only be avoided by proper use of randomization in selecting a sample. We wish to use a sample selection scheme which can be repeated by anyone who understands the technique being employed, but the units selected should in no way depend on the individual making the selection. It is considered good practice to record the steps in selection so they may be verified to insure the technique is being carried out properly. The selection schemes discussed can be used with any survey design, but the selection methods are most frequently used to achieve one or all of the following objectives: (1) greater efficiency in terms of a smaller sampling error, (2) ease of summarizing the data prior to making estimates, or (3) to insure proper execution of the sampling procedure by personnel.

3.2 Simple Random Sampling

This implies that the probabilities of selection for all sets of sample units of size n are equal and each unit selected from the frame will have the same expansion factor N.

3.2.1 Sampling With Replacement

The universe contains N units and we wish to select n units for the sample. Let us suppose N = 850, then we develop a rule so that one and only one integer between 1 and 850 is associated with a particular unit in the frame. To achieve the selection of n units, we choose n three-digit numbers from a random number table between 1 and 850. If we encounter a number greater than 850, we ignore it because none of our units have been labeled with numbers greater than 850. If the same random number is selected more than once, then the unit identified by this number must be included in our summary or analysis once for each time the random number is selected. We illustrate this for n = 10, using the

random-number sheet attached at the end of this chapter. To guard against different people (or the same person) always using the same set of random numbers, we find a new random starting point in the table each time a sample or replicate is selected by visualizing the table as consisting of a given number of pages of C columns (25 per page) of three-digit numbers from 001 to 850 and R rows (40 per page) of three-digit numbers. We enter the table by finding a starting column and starting row on a particular page. One procedure by which we select a starting point in the table of random numbers is as follows: On page one of the random-number table, the first one-digit number was used to select the page, the second two-digit number selected the starting column on the page, and the third two-digit number selected the starting row. (In case any of the numbers exceeds the number of pages, columns or rows, then we use the first number equal to or less than the maximum pages, columns or rows.) We proceed from the starting point down the column until we have 10 numbers. If we do not obtain 10 random numbers before we exhaust all the numbers in the starting column, we proceed to the top (or bottom) of the next column to the right (or left) until we obtain the 10 numbers.

The following two samples of size 10 were selected by this procedure. The starting points were: (1) page 2, column 08, row 40, and (2) page 2, column 21, row 04.

	Sample Number	
	1	2
Random Number Selected	551	219
	479	667
	284	698
	094	024
	307	211
	417	166
	654	502
	749	518
	760	027
	622	364

Since a random number (or three-digit integer) can occur any number of times in our table, the probability of selection on any draw is $\frac{1}{N}$, and the probability of not being selected $\frac{N-1}{N}$. Thus, the probability of selecting each unit is

constant, hence the probability of selecting the n units is

$\frac{1}{N} + \frac{1}{N} + \ldots + \frac{1}{N} = \frac{n}{N} = \frac{10}{850}$, or the expansion factor for the 10 units

($\underbrace{\qquad\qquad}_{n \text{ terms}}$)

is 85, and for an individual unit, 850.

3.2.2 Sampling Without Replacement

The labeling of the universe units is the same as in 3.2.1. To achieve the selection without replacement, we insist on n distinct random numbers between 1 and N. We proceed as in 3.2.1 except if we encounter a number previously selected, we do not use it (or cross this number out) and simply proceed to select the next number. The following two samples of size 10 were selected using this technique. The starting points were: (1) page 2, column 04, row 32, and (2) page 4, column 10, row 19.

	Sample Number	
	1	2
Random Number Selected	543	574
	241	360
	442	446
	814	358
	170	027
	601	026
	315	039
	841	188
	110	415
	488	456

Since a random number cannot occur in our sample more than once, a universe unit cannot be selected more than once. The probabilities of selecting a unit from the units available on the successive 10 selections are:

$$\frac{1}{N}, \frac{1}{N-1}, \frac{1}{N-2}, \frac{1}{N-3}, \frac{1}{N-4}, \frac{1}{N-5}, \frac{1}{N-6}, \frac{1}{N-7}, \frac{1}{N-8}, \frac{1}{N-9}$$

and the probabilities of nonselection are:

$$\frac{N-1}{N}, \frac{N-2}{N-1}, \frac{N-3}{N-2}, \frac{N-4}{N-3}, \frac{N-5}{N-4}, \frac{N-6}{N-5}, \frac{N-7}{N-6}, \frac{N-8}{N-7}, \frac{N-9}{N-8}, \frac{N-10}{N-9}.$$

The probability of selection for any of the n units is $\frac{1}{N}$ and of nonselection $\frac{N-1}{N}$. However, this may not be immediately obvious. Consider the probability of obtaining the i^{th} unit on any of the 10 possible selections.

$$P(i) = P(1^{st} \text{ draw}) = P(i) = \frac{1}{N}$$

$$P(i) = P(2^{nd} \text{ draw}) = P(\bar{i}\,1)\,P(i|2) = \frac{N-1}{N} \cdot \frac{1}{N-1} = \frac{1}{N}$$

$$P(i) = P(3^{rd} \text{ draw}) = P(\bar{i}\,1)\,P(\bar{i}|2)\,P(i|3) = \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \frac{1}{N-2} = \frac{1}{N}$$

$$\vdots$$

$$P(i) = P(10^{th} \text{ draw}) = \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \frac{N-3}{N-2} \cdot \frac{N-4}{N-3} \cdot \frac{N-5}{N-4} \cdot \frac{N-6}{N-5} \cdot \frac{N-7}{N-6} \cdot \frac{N-8}{N-7} \cdot \frac{N-9}{N-8} \cdot \frac{1}{N-9} = \frac{1}{N}$$

Consequently, the probability of selecting n units is: $\underbrace{\frac{1}{N} + \frac{1}{N} + \frac{1}{N} + \ldots + \frac{1}{N}}_{\text{n terms}} = \frac{n}{N}$

The expansion factors for a sample of size 10 are the same as given earlier: 85 for the set of 10 units and 850 for an individual unit.

3.3 Random Systematic Sampling

The extreme popularity of systematic sampling is due to convenience in use and gains in efficiency which may result from taking advantage of natural stratification which may exist in the way the frame was assembled. In fact, the frame should be assembled (or reassembled) in such a way as to create the most efficient possible ordering of the units so the sampling intervals will roughly correspond to some kind of strata. However, the sampling intervals should not be viewed as strata unless provision is made to obtain unbiased estimates of the sampling error. Rather, it is proper to view a single systematic sample based on one random start as a special kind of cluster. It should be noted that in using a single random number to start the selection process results in only one cluster of units no matter how many times the sampling interval is added to the random start. Consequently, it is not possible to obtain an unbiased estimate of the sampling error.

3.3.1 Unclustered Units in the Frame

In this case, it was possible to directly identify each of the N units in the universe with a selected random number. Commonly, places or names of people

may be assembled in either an alphabetical order or by political units (or geographic areas). It is seldom clear for an alphabetical ordering what kind, if any, natural stratification may exist, but some type of stratification may exist even though it cannot be identified. As a result, it is frequently assumed that the ordering is equivalent to a random ordering for purposes of approximating the sampling errors by treating the cluster as a simple random sample of n units. Units listed by political or geographic order can frequently be rearranged to create a stratification of the units which will increase the efficiency. Or if information on the magnitude of some characteristic which is related to the content of the survey is available, a more efficient stratification will generally result.

We will consider a small universe to help illustrate the nature of the clusters which are created by systematic sampling. Generally, the clusters formed are much more efficient than the compact or contiguous units which could be used to form natural clusters. The clusters formed in systematic sampling are composed of units which may be scattered over the entire universe and hence these units are less apt to be alike. This results in the variability within clusters being increased while the variability between clusters is decreased. When all units in a cluster are sampled, no contribution to the sampling error is made by the variability within clusters. If N = 150 and we are interested in samples of size 10, the universe will consist of fifteen clusters. The sampling interval (i) will be 150 ÷ 10 = 15 and a random number between 1 and 15 will be chosen to select a sample of 10 elementary units, or one cluster.

Units Selected By R.N.	Random Numbers (i.e., clusters)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Same as R.N.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R.N. + i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
R.N. + 2i	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
R.N. + 3i	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
R.N. + 4i	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
R.N. + 5i	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
R.N. + 6i	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
R.N. + 7i	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
R.N. + 8i	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135
R.N. + 9i	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150

Clearly, the selection of only one cluster does not permit calculation of the sampling error. In order to insure an unbiased estimate of the sampling error in systematic sampling, two or more clusters must be selected. Generally, the selection of two or more clusters of the type in our table are referred to as replicated or interpenetrating samples, because each cluster of 10 units is spread over the universe in the same way. The use of replicated samples is required for estimation of sampling errors and frequently as a means of evaluating nonsampling errors.

3.3.2 Sampling from the Dominican Republic Frame

We will consider several examples using simple random sampling and systematic sampling techniques for the Province of Salcedo for which the class has constructed a sampling frame with five land use strata.

3.3.2.1 Selecting a Simple Random Sample

A. In sample selection, the first step is to decide on the size of the average segment (the ultimate sampling unit selected for enumeration). In the Dominican Republic, the sizes were as follows:

Stratum	Size
I, II	2 km^2
III, IV	4 km^2
V	1/2 km^2

All cities, no matter how small, were considered to have at least one sampling unit.

B. Sampling units need to be assigned to each count unit. Divide the area of the count unit by the expected segment size and round the quotient to the nearest whole number to get the number of sampling units for each count unit. List the sampling units for each count unit. List the sampling units on the count unit identification sheet in the column marked "S.U." Accumulate a total of the sampling units for the province in the column marked "CUM."

C. A simple random sample of segments is to be selected from Stratum I. The starting point in the random number table is: page 3, column 5, row 24.

To select a segment:

1. From the starting point, pick a random number from 1 to N_h where N_h is the total number of sampling units in the particular stratum.

2. Determine in which count unit the selected sampling unit falls. On a segment location sheet, record the segment number, the stratum number, the count unit number, and the number of sampling units in the count unit. The final column may be used to record the name of the city for segments in Stratum V or any other pertinent information.

3. Find the selected count unit on the map and divide it into the assigned number of sampling units using the best available boundaries. Each sampling unit need not be exactly equal to the expected size. Size may vary within the tolerance range stated below (commonly ± 50%):

Stratum	Tolerance
I, II	1-3 km^2
III, IV	3-5 km^2
V	1/4-3/4 km^2

4. Number the sampling units in the count unit beginning in the northeast corner and proceeding in a serpentine fashion as before. Select one at random and identify it with the segment number in red on the map.

We will only consider selection from Stratum I. Each of you might have arrived at a different number of sampling units (or possibly even different strata) for Salcedo Province. However, you are to use the frame you (or your team) constructed. The original frame construction is shown in the sketch in training set #1 with the following number of sampling units: Stratum I - 89 S.U., Stratum II - 115 S.U., Stratum III - 4 S.U., Stratum IV - 2 S.U., and Stratum V - 3 S.U. The example which follows was based on this frame. However, you will need to use the numbers for the frame you constructed in the exercise.

A simple random sample of 5 segments is to be selected without replacement from the 89 S.U. in Stratum I. The starting point in the random number table was: page 3, column 5, row 25. The units selected were: 04, 34, 72, 30, and 44. These are shown by the circles on the sketch.

3.3.2.2 Selecting Replicated Systematic Samples

Next, we consider a systematic sample of 5 units to illustrate the nature of the "paper" substrata which can be created by the choice of the number of substrata or sampling interval (see section 4.2.4 for discussion). The purpose of this is to illustrate two points: (1) That a sampling frame can be modified, either to create additional stratification or to modify the number (or size) of sampling units; and (2) to show one type of paper substrata based on geographic proximity. If we wished to achieve the maximum degree of natural stratification, we would have 5 "paper" strata and the sampling interval would be 89/5 = 17.8. To avoid fractional sampling intervals and to create 5 paper strata with equal numbers of sampling units, we divided the largest unit into two units or the largest two adjoining units into three units. This results in the strata now containing 90 S.U., that is, 5 substrata of 18 units or a sampling interval of

90/5 = 18. The natural boundaries of these paper strata are shown by the dashed lines and with letters a, b, c, d, and e. A systematic sample of 5 S.U.'s is selected, or one cluster out of the 18 possible clusters is selected. Using a random starting point on page 3, column 20, and row 13, the random number obtained between 1 and 18 is 8, which results in units numbered 8, 26, 44, 62, 80 being selected. These units are shown by the X's on the sketch.

However, the usefulness of the stratification or the sampling error of the systematic selection can only be shown if two or more replicated samples of the same type are selected. This is normally a part of the postsurvey analysis. We have looked at the creation of "paper strata" within the land use (or primary) strata. The structuring of the sampling units within a primary stratum to create "paper strata" is done (1) to insure a more representative distribution (increased efficiency) of the sample as compared with simple random sampling over the province, (2) to use a simple and convenient method of selecting a systematic sample, (3) to provide a basis for measuring the sampling error of a systematic sample, (4) to provide a postsurvey basis for analysis which may suggest a more efficient survey design for subsequent surveys, (5) to enable assignment of interviewers to replicates to reduce the influence of the interviewer on survey results.

It should be pointed out that while the nature of the frame modification was very minor in this example (from 89 to 90 S.U.) there may arise situations in which the modification may be more substantial. If, for instance, the characteristic of interest occurs infrequently, it may be desirable to increase the sample unit size, say from 2 km^2 to 4 km^2, so most of the sampling units will have present the characteristic which we are interested.

3.3.3 Exercise 5:

Using the sampling frame constructed for Salcédo Province, each student is to select two replicated systematic samples of size 5 in Stratum I.

3.3.4 Sampling from a List Frame

Several exercises in sampling from the list frame assembled in chapter 2 will be undertaken by each student. First, a sample of primary units (clusters) based on selecting the first-stage units with probability proportional to the number of second-stage units, then selecting a constant or fixed number of

secondary units for each primary unit. The second exercise will consist of selecting replicated samples from each of the strata which were created in chapter 2.

3.3.4.1 Sample Selection Using a Two-Stage Process

Twenty primary units are to be selected using unequal probabilities of selection with replacement based on the number of name-address units in each city. In the area frame, the count units would be analogous to the cities and the segments would be analogous to the name-address units. For the second stage in the selection, choose two name-address units using simple random sampling without replacement. This probability of selection scheme results in each name-address unit having the same probability of selection. When this feature is present, the sample is frequently called a self-weighting design. The equal probability of selection for each name-address unit results because:

(1) The probability of selecting an individual primary unit is:

$$P(i) = \frac{M_i}{\sum_{i=1}^{C} M_i}$$

where M_i is the number of name-address units in the i^{th} primary unit and

$$\sum_{i=1}^{C} M_i = N$$

is the total number of name-address units in the C primary units in the frame.

(2) The probability of selecting a secondary unit within a given primary unit is:

$$P(j|i) = \frac{1}{M_i}$$

Or, the unconditional probability of selecting any name-address unit is

$$P(ij) = P(i) \cdot P(j|i) = \frac{1}{\sum_{i=1}^{C} M_i} = \frac{1}{N}.$$

3.3.5 Exercise 6:

Using the strata constructed for the list frame in chapter 2, each student is to select a simple random sample of five name-address units in each stratum.

RANDOM SAMPLE NUMBERS - Page 1

↓Page ↓Col. ↓Row				
93049	79367	00812	41365	04515
62865	09576	97207	33739	78345
00800	72496	24767	61768	07228
64340	02224	48336	14891	72188
92168	52692	31224	12185	43065
20494	18813	16242	40257	66402
87693	30242	70545	69128	51528
05567	05561	82071	07234	67690
85166	37189	75671	33879	27411
26704	47922	56650	40236	66207
01047	81624	77395	62310	41501
58183	21952	84098	28913	55736
64667	57092	21315	04731	71877
27149	13843	09817	09407	88276
66232	80293	74502	36925	60184
40500	21406	00571	87320	81683
35892	49668	83991	72088	30210
54819	26094	51409	21485	94764
64224	47909	09994	23750	17351
36913	58173	45709	83679	82617
64254	64745	10614	86371	43244
82018	25536	74031	31807	70133
28833	44043	96215	21270	59427
96879	27659	95463	53847	40921
95938	76014	99818	16606	19713
97154	71237	06073	57343	51428
78790	17026	59008	28543	11576
25034	59325	08844	95774	49323
70116	44091	88505	15575	44927
66904	23000	73259	68626	98962
91171	28299	62619	81550	46798
74547	13260	79262	55831	83784
30448	14154	75795	39465	82353
06584	29867	45898	66415	89349
68548	86576	14344	75889	04514
49319	50206	22024	56124	50749
81034	86779	34622	70859	33045
68905	44234	18244	31602	38388
88530	72096	44459	31449	93182
37227	11302	04667	32526	64712

04808	99531	47991	46064	80467
71924	64882	94893	82935	99076
56410	89552	28404	74525	74212
38851	16144	99542	27481	21992
91428	10589	09454	43308	66753
40083	17141	30702	31997	69856
93419	10474	41796	88285	02448
03704	65516	65448	20203	21189
78181	90060	74904	42627	16638
45972	93572	76011	03426	50226
60898	63968	62264	64603	51866
40398	54180	65869	87977	02799
68245	76912	01222	59516	36438
27019	15248	66444	25267	05171
99868	88894	43769	52239	05919
87904	74135	53842	59520	23979
68851	41049	97190	53984	04773
71742	57223	66599	86071	01901
02742	48803	17823	22093	43907
56181	96052	67211	61712	54590
55355	61548	55988	47309	23749
78961	41072	09876	18903	30292
92654	97226	53434	77025	63892
13757	37719	84450	02697	60309
05776	85945	74651	00216	50842
71039	83083	60427	78495	99809
61672	01184	46438	27698	40652
42988	77983	58708	42176	67356
13652	16640	27896	26907	86760
53186	97859	97213	19859	41037
47890	10690	26486	38744	25943
65654	34629	88831	97253	67282
00324	17120	39900	67135	42772
48244	26191	88421	90491	83290
64081	47704	15018	45600	17241
60617	06414	56596	63011	24193
72860	18452	42983	23931	11789
04631	55283	19605	34163	86540
06884	15444	09310	17048	24243
26611	09551	82626	38194	58432

Starting point

Starting point

Starting point

22224	02627	91576	16781	89184
24473	42096	76920	88864	54164
38582	21871	14672	93362	67981
46094	43845	91838	79574	08003
91061	31674	73729	99315	16699
00397	56753	53158	71872	68153
14328	44708	72952	27048	67887
88534	87112	68614	83073	88794
97347	87316	73087	77135	71883
01366	72976	01868	51667	63279
37106	20523	21584	93712	83654
06476	70603	97122	44978	78028
81717	48410	94516	15427	85323
51583	69788	41758	55004	30992
50120	33884	83655	88345	69602
89761	23053	77480	28683	68324
88943	66660	11057	98849	29499
71685	97247	79368	43710	80365
17402	66300	94385	01717	96191
52606	39860	92127	42588	93307
66035	07223	76264	29148	68652
21565	30786	45403	33782	93424
88735	75275	03080	77653	55430
50404 Starting	80166	28017	52611	60012
80834 point	11317	93109	91857	47904
26872	72927	79021	51571	68825
16530	96086	17329	87959	23727
84644	00448	86828	50552	84832
88620	72894	94716	84622	49771
22209	78590	68615	58113	23727
04795	53971	14592	39634	03855
54291	56045	61635	32186	86651
30654	48543	18339	65024	33386
11123	08732	49393	12911	75803
56577	51257	83291	12329	17827
58987	02026	42969	59144	84349
16851	99197	70476	77113	46320
02104	49435	77706	18924	24957
54440	07893	31618	35707	65130
87681	42543	69847	81848	32034
24337	61634	52574	83649	28725
62557	25292	72781	17186	10393
02913	03885	58822	82941	43806
68706	87619	13846	56197	27151
05930	33213	78416	00194	91369

40083	17141	30702	31997	69856
93419	10474	41796	88285	02448
03704	65516	65448	20203	21189
78181	90060	74904	42627	16638
45972	93572	76011	03426	50226
60898	63968	62264	64603	51866
40398	54180	65869	87977	02799
68245	76912	01222	59516	36438
27019	15248	66444	25267	05171
99868	88894	43769	52239	05919
87904	74135	53842	59520	23979
68851	41049	97190	53984	04773
71742	57223	66599	86071	01901
02742	48803	17823	22093	43907
56181	96052	67211	61712	54590
55355	61548	55988	47309	23749
78961	41072	09876	18903	30292
92654	97226 Starting	53434	77025	63892
13757	37719	84450	02697	60309
05776	85945 point	74651	00216	50842
71039	83083	60427	78495	99809
61672	01184	46438	27698	40652
42988	77983	58708	42176	67356
13652	16640	27896	26907	86760
53186	97859	97213	19859	41037
47890	10690	26486	38744	25943
65654	34629	88831	92753	67282
00324	17120	39900	67135	42772
48244	26191	88421	90491	83290
64081	47704	15018	45600	17241
60617	06414	56596	63011	24193
72860	18452	42983	23931	11789
04631	55283	19605	34163	86540
06884	15444	09310	17048	24243
26611	09551	82626	38194	58432
04808	99531	47991	46064	80467
71924	64882	94893	82935	99076
56410	89552	28404	74525	74212
38851	16144	99542	27481	21992
91428	10589	09454	43308	66753

CHAPTER 4 - EXAMPLES OF AGRICULTURAL SURVEYS

4.1 Introduction

In this chapter some schemes and techniques previously discussed will be presented to deal with "typical" problems of agricultural surveys. This will be done primarily through the discussion of two actual surveys and an analysis of the survey data. The estimates and their sampling errors will be presented, based on making the computation with a desk-top calculator. The students will also code and punch one of the data sets for analysis with the computer. The listing of the computer program is included in the appendix along with the necessary input cards for one of the two surveys. However, no lengthy discussion of missing, incomplete, or the editing of data will be attempted in conjunction with the examples, but rather it will be assumed this function has been completed for each sampling unit prior to preparing the estimates.

4.2 Some Design Considerations

The difference between a statistical investigation and one not statistical is that with the statistical investigation, we are able, in the end, to evaluate the main sources of uncertainty in the results. The more we know about what is wrong with a result, the more useful it becomes. It is possible we may learn more from what went wrong in a study than from what went right. Ability to evaluate the uncertainties in a result is not an accident. Only by use of appropriate statistical design, which will include statistical controls for the detection of blemishes and blunders in the measurements, interviews, and processing, is it possible to evaluate the possible effects on the results that arise from the chief sources of nonsampling errors.

Use of a nonprobability sample (one where some element of human judgment rather than random numbers enter directly into the selection of the sample units) is worth no more than the reputation of the man who prepares the study. The reason is that there is no way except by his judgment to set limits on the margin of uncertainty of the estimate. Probability samples offer relief from the uncertainty of the magnitude of sampling variation in a result. "Judgment samples" ignore the problems of nonresponse and the biases introduced by it, but do eliminate occasional high cost and inconvenience of travel.

The first question to ask when a statistical investigation is proposed is: Can you get from any unit in the universe (person, household, farm, business

establishment, piece of property), no matter how you select it, the information that you require? If the answer is yes, then ask a second question: If you were to elicit the desired information from every unit in the universe, would the compilation of results be useful? If the answer is yes to the second question, then a statistical investigation might be worth consideration. If the answer to either question is no, then take a second look: revise the aims, or the method, or abandon the survey.

Some surveys are particularly difficult to design for efficiency because of high costs for doing certain required operations. These are essentially of three types: those with (1) high costs of seeking out and identifying the elements in the universe to be surveyed; sometimes these are referred to as rare cases, (2) high costs of making observations or collecting the desired information, and (3) the case where the planner is quite uncertain what the costs of the various operations might be. Each type is considered briefly.

4.2.1 High Cost of Identifying Elements for Special Groups

Surveys directed at certain special groups may be particularly difficult to deal with because of high costs and other problems. A special group may be "dairy farms," "cotton farms," land owners with timber for sale, households with unemployed persons, dealers in fertilizers, persons with certain diseases, etc. Generally, the special group is a subgroup of some larger and generally more "framable" or accessible group like: households, farms, or business establishments, etc. The elements in the subgroup are not identified prior to the proposed survey; it is part of the burden of the survey to seek out and identify the subgroup membership. This search procedure, frequently called screening, if carefully carried out can be expensive and leave very little of the budget for obtaining observations on the subgroup itself. To deal with the problem of identifying rare or special-group elements, sampling statisticians usually follow several procedures: (1) employ all the appropriate and generally known principles of sampling, such as: choice of sampling unit, increasing unit size to obtain a nearly constant expected number of rare elements per sampling unit, use of measure-of-size information, stratification, double sampling, etc., (2) get hitherto unknown information on lists, estimates of total, clustering pattern of the group elements, innovation of more efficient search techniques, etc., and (3) joint use of more than one sampling frame.

4.2.2 High Cost of Information per Unit

In many rural communities, growers do not know the actual area planted to crops, but do know how much seed, or how long (in days or hours) it took them to plant the area. In such situations, it may be necessary to subsample every k^{th} planting and measure the field size with special workers using rules, tapes, or other equipment to determine accurately the area planted. In sampling areas planted to corn, wheat, rice, or coffee for yield per acre, the cost of determining yield per acre by some objective means (as crop cutting) may be very expensive. It may be possible to get the grower's estimate of the yield for a field (or farm) and then count heads (or beans) in the same field and take sample heads to a central point for weighing or threshing. If the grower's yield estimates are correlated with the more expensive determination of the plant characteristic per unit area, this type of double sampling can be very effective in reducing costs and providing yield data rigorously defined in terms of weight per unit.

Obtaining data on crop input and other farm expenditures may require compiling records or very lengthy interviews to verify actual costs. It may be a better use of survey resources to merely determine the different kinds or categories of costs incurred for each farm in a large sample. Then in a smaller sample, obtain detailed dollar costs per acre (or per farm) for the crops or farm characteristics of interest. Such procedures involve two-stage sampling to obtain the data which is too costly to secure for all farms because of the length of interview, availability of respondent data, or special training of workers to secure useful data.

4.2.3 Unknown Operational Costs

The cost of performing each of the operations in a survey and of feasible alternative procedures is seldom known with suitable accuracy, particularly in the large-scale nonrepetitive surveys. In such cases, the planners frequently need more information on costs and other operational matters, such as clarification of concepts and definitions involved in data collection. While pilot surveys cost money and time, which may reduce funds for the survey proper, they can be viewed as a two-step survey that can lead to more effective use of available funds. Pilot surveys are also an excellent training device and are useful for "debugging" procedures. Since many surveys are repeated at various points in time (for example, quarterly), the first survey can be viewed as a pilot or smaller survey which will lead to more efficient surveys in the future. Where

possible, it may be practical to conduct the pilot survey on fruit or tree crops (which remain constant over time) for the first and smaller survey. Also, the survey planner will need to consider whether to use the same sample repeatedly or to use independently selected samples for each survey. In general, a mixture is desirable for most purposes, with approximately one-half of the sample being changed for each survey.

4.2.4 Use of Paper Strata

A technique for creating substrata which are referred to as "paper strata" can be used to increase sampling efficiency without requiring detailed information on auxiliary variables or on variances which is required for detailed stratification and optimum sample allocation. It is discussed at this point in the course, since the paper strata are designed to satisfy the needs of the data collection system. Additionally, they make it possible to employ simple sample selection procedures and summarization methods in calculating the estimates and their sampling errors, using equal probabilities of selection.

The paper strata are assembled from the existing frame units to reflect similar crop types (or types of farming) as well as microgeographic stratification within the primary land use strata. Most of the information used for this task is subjective in nature, being based on crop types by localities, census data, elevation contours and houses on topographic maps, and other local sources of information useful for agricultural surveys. Using this information, a detailed stratification within the primary strata is achieved. If the purpose of the data collection system is to obtain household and family living information, the paper strata might be based on factors such as value or type of housing, personal income, or ethnic background as well as microgeographic considerations. This type of stratification might be appropriate even for an agricultural system within an urban land use stratum.

This technique may be used with a previously finished frame or a preliminary listing of units for a new frame being constructed. It is achieved by reassembling the listing of primary frame units (and occasionally subdividing or modifying primary frame units) within a land use stratum in a new order so that "like areas" are listed consecutively. Since the stratification is achieved by a reassembly of existing listing forms, they are called paper strata. It is convenient to assign the same total number of sampling units to each paper stratum and to select the same number of sampling units in each

paper stratum within a primary land use stratum. These two features make it easy to obtain survey tabulations or additional tabulations which may be requested for geographic areas.

The number of paper strata within a land use stratum is usually the same as the number of sampling units per replication in an interpenetrating design. Thus, the sample size (n_h) for a land use stratum is a product of the number of replications (m_h) and the number of paper strata (k_h); that is, $n_h = m_h k_h$. Since it may be necessary at a later time to reduce the sample size as the precision requirements are changed, it may be desirable to be able to reduce the number of paper strata, (i.e., sampling units per replication). Consequently, it is convenient to let the number of paper strata be a multiple of 2 or 4 so that pairs or groups of four paper strata can be combined if a smaller sample needs to be allocated to the land use stratum. Of course, the sample size can also be changed by increasing or decreasing the number of replications. This type of substratification and flexibility in setting number and size of paper strata is believed to be well suited to the needs of countries making a major modification in or developing a new data collection system.

4.3 Coffee Survey in the Dominican Republic

4.3.1 Background and Design

In 1970 the Secretariat of Agriculture was seeking technical assistance to implement a system of sampling. At about the same time, the International Coffee Organization in London was seeking to update information on the Dominican Republic's coffee industry. The urgent need of the ICO for special information and the domestic needs for other regular agricultural data provided sufficient demand to construct an area sampling frame and conduct a limited agricultural survey. In March 1971, through the efforts of AID, the area sampling frame construction was started by SRS personnel in Washington, D. C. under a PASA agreement, using available cartographic materials and aerial photographs. In January 1972 the completed materials and a sample of 160 area segments selected for the first survey on coffee were shipped to the Secretariat of Agriculture. In addition, the Statistical Reporting Service was asked by AID to provide technical assistance in training, conducting, and coordinating a system of agricultural data collection through preparation and interpretation of estimates. An SRS individual was assigned to assist in this task for a two-year period.

The sampling plan for the first coffee and cacao survey was a stratified replicated (or interpenetrating) design. Two primary strata were the Intensive Coffee and Cacao Stratum and the group of the remaining four land use strata in the area frame. Within each of these primary strata, 20 equal-size paper strata were created in Stratum II and 10 equal-size paper strata in the combined Strata I, II, III, IV, V. The paper strata were developed by listing the frame units by geographic areas within the provinces and across provinces. It was generally possible to develop similar land use areas across boundaries by making several alternative or preliminary listings of the frame units prior to adopting the final paper strata. In addition, five geographic areas were identifiable in the two primary strata. Within each stratum replicated systematic samples of size 20 and 10 were selected, starting with a random number in the first listed paper stratum and using nonreplacement sampling for successive samples of size 20 and 10. Each selected replicate constituted a cluster of size 20 (or 10) for analysis purposes, since each sampling unit in the frame is assigned to one and only one cluster of size 20 (or 10) within each stratum. Each cluster so chosen contains one segment from each paper stratum which insures representativeness in terms of the geographic and crop type pattern present in the frame as constituted.

The sampling unit size was based on experience in the United States with modification to accommodate the availability of natural boundaries and expected number of farms per unit area. In the intensely or moderately cultivated areas ample natural boundaries appeared to be available and the average size of holding was expected to be small. Consequently, a sampling unit which would average two square kilometers was chosen. At a later time each sampling unit can be split in half if either the postsurvey analysis or cost considerations indicate a smaller unit of one square kilometer would be desirable. In the populated places a sampling unit of one quarter of a square kilometer was used and in the extensive marginal agricultural area a unit of four square kilometers was used. Obviously, smaller sampling units could be obtained by subdivision at a later date. Consequently, the size of the sampling unit (i.e., segment) in Stratum II was two square kilometers and in the group of strata (I, III, IV, V) was either 1/4, 2, or 4 square kilometers, depending on which of the four land use strata the sampling unit was selected from. However, each replicate in the group of strata (I, III, IV, V) was expected to contain the same proportion of 1/4, 2, or 4 square kilometer units. The use of the five strata for later surveys was

recommended to permit maximum benefits from stratification and minimum variation in sampling unit size within strata.

The Sampling Frame Configuration and Sample Size for the Initial Coffee Survey

Strata	Expected S.U. size		Total number S.U. N_h	Sample allocation (no. S.U.) $= n_h$	Number of replicated samples $= m_h$	Number of paper strata (cluster size) $= k_h$
	Frame	Coffee survey				
II	2 km^2	2 km^2	978	120	6	20
I	2 km^2		3,193			
III	4 km^2	3.4 km^2	7,873	40	4	10
IV	4 km^2		2,784			
V	.5 km^2		261			
	3.3 km^2		15,089	160		

Each segment within a stratum was selected with the same probability, which makes it possible to do most of the summarization without expanding the data. While the sample size for the noncoffee stratum (I, III, IV, plus V) was believed to be smaller than desirable, it was anticipated that a much larger general-purpose survey would be conducted later in the same year. Consequently, a second estimate for these strata would be available and could be expected to have a smaller sampling error. The basis for estimating the total tareas of coffee and its sampling error squared is given below.

Let

x_{hij} = the total tareas[1/] of coffee reported in the j^{th} segment (S.U.) in the i^{th} replicate of the h^{th} stratum (values given in table 3)

1/ 15 tareas = approximately 1 hectare

$x_{hi\cdot}$ = total tareas of coffee reported in the i^{th} replicate of the h^{th} stratum (column totals in table 3)

$\hat{T}_{hi}$ = estimated total tareas of coffee for stratum based on i^{th} replicate

$$= \frac{N_h}{k_h} \sum_{j=1}^{k_h} x_{hij} = \frac{N_h}{k_h} x_{hi\cdot}$$

$\hat{\bar{T}}_h$ = estimated average total tareas of coffee for the h^{th} stratum

$$= \frac{1}{m_h} \sum_{i=1}^{m_h} \hat{T}_{hi}$$

N_h = total number of segments (S.U.) in stratum (subpopulation)

k_h = number of segments in cluster in h^{th} stratum

m_h = number of clusters in h^{th} stratum

n_h = $m_h \cdot k_h$ = number of segments in stratum

and the variance of the estimated total is

$$V(\hat{\bar{T}}_h) = \sum_{i=1}^{m_h} \frac{(\hat{T}_{hi} - \hat{\bar{T}}_h)^2}{m_h(m_h - 1)} \quad .$$

4.3.2 Calculating Estimates and Sampling Errors

A summary of the reported tareas in coffee by segment for three of the replicates in Stratum II and two replicates in the other stratum are presented in table 3. The estimate of the total tareas planted to coffee and its sampling error can be derived, based on two alternative methods of summarizing the survey data. The two methods produce identical results. The first method is based on working with unexpanded data by segment, cluster, and stratum given in table 4. The computations with the expanded data are left as an exercise for the student. The estimate and its sampling error for Stratum II are given below for systematically selected replicated samples.

The variance is:

$$V(X_{hi}) = \frac{17976^2 + 20799^2 + 22994^2 - (61{,}769^2 \div 3)}{2} = 6{,}327{,}946$$

and variance for average replicate

$$V(\bar{X}_h) = \frac{V(X_{hi})}{3} = 2{,}109{,}315$$

where X_{hi} = the reported total tareas for the replicate (i.e., cluster)

$\bar{X}_h$ = the average reported total tareas for all replicates in Stratum II

= 20,589.67

and the coefficient of variation of $\bar{X}_h$ is

$$CV(\bar{X}_h) = \frac{\sqrt{2,109,315}}{20,589.67} = .07054 \text{ or } 7.05\%$$

or, using table 1, the expanded replicate total $T_{hi} = (48.9)X_{hi}$.

Hence, the estimated total tareas of coffee for the stratum is:

$$\bar{T}_h = (48.9)\bar{X}_h = 1,006,835.$$

The variance of T_{hi} is

$$V(T_{hi}) = (48.9)^2 \, V(X_{hi}) = 15,131,447,755$$

and $$V(\bar{T}_h) = \frac{V(T_{hi})}{3} = (48.9)^2 \, . \, V(\bar{X}_h) = 5,043,815,121 \, .$$

Therefore

$$CV(\bar{T}_h) = \frac{\sqrt{5,043,815,121}}{1,006,835} = .0705 \text{ or } 7.05\% \, .$$

4.3.3 Postsurvey Analysis

Although the survey error has been derived based on the sampling procedure used, the question of whether there was either a more efficient or equally efficient method of selecting the sample can also be answered for several alternative schemes. Consider three methods of selecting a sample of 60 segments from the 978 in the stratum: (1) A sample stratified by paper strata; that is, a simple random sample of 3 from each of the 20 paper strata, (2) a simple random sample of 60 without any paper strata, and (3) the three replicated samples using the paper strata to provide or control geographic dispersion of each replicate. This estimate and its sampling error were computed in 4.3.2.

Method 1: The postsurvey analysis of expected sampling errors is based on the selection in each paper stratum of independent random samples of size m. A variance is computed for each paper stratum (rows in tables) from the m segments and "pooled" (or averaged) across all paper strata. (See table 5.)

$$V(\bar{X}_{h.}) = \sum_{j=1}^{20} \frac{V(X_{hj})}{m_{hj}} = \frac{\sum_{j=1}^{20} V(X_{hj})}{3} = \frac{10,355,251}{3} = 3,451,750.3$$

and the coefficient of variation based on this stratified method of sampling is:

$$CV(\bar{X}_h) = \frac{\sqrt{3,451,750}}{20,589.6} = .0902 \text{ or } 9.02\%$$

which is slightly less efficient than the systematically selected replicates. Or, based on the expanded population totals by paper strata, the estimate and its variance are:

$$\bar{T}_h = N \bar{X}_h = (978)(1029.48) = 1,006,831$$

$$V(\bar{T}_h) = (978)^2 V(\bar{X}_h) = 8,253,854,335$$

and

$$CV(\bar{T}_h) = \frac{\sqrt{8,253,854,335}}{1,006,831} = .0902 \text{ or } 9.02\% .$$

Method 2: The postsurvey analysis of expected sampling error is based on selecting a single simple random sample of size 60 (i.e., $m_1 k_i = n_1$). A result given by Cochran, 2nd Edition, page 139 is used to derive the variance for a frame without the paper strata.

$$V(x) = \sum_{j=1}^{k} W_k S_k^2 - \sum_{j=1}^{k} \frac{W_k S_k^2}{m_k} + \sum_{j=1}^{k} \frac{W_k^2 S_k^2}{m_k} + \sum_{j=1}^{k} W_k \bar{x}_k^2 - \left(\sum_{j=1}^{k} W_k \bar{x}_k\right)^2$$

where k = 1, 2 ... 20 (number of paper strata), and n_k = sample size in a stratum.

$W_k = \frac{N_k}{N} = \frac{1}{20}$; since all paper strata are the same size

$$N_1 = N_2 = \ldots = N_k .$$

S_k^2 = variance within paper strata (see col. 3, table 5).

$$V(x) = 517,762 - \frac{517,762}{3} + \frac{517,762}{60} + \frac{34,708,000}{20} - \frac{423,931,628}{400} = 1,020,746$$

$$V(\bar{x}) = \frac{V(x)}{n_h} = \frac{1,020,746}{60} = 17,012.4 \text{ (for segment mean).}$$

The population total for a single random sample in Stratum I is:

$$\bar{T} = N_1\bar{x} = (978)(1{,}029.48) = 1{,}006{,}831$$

and

$$V(\bar{T}) = N_1^2\, V(\bar{x}) = (978)^2\, 17{,}012.4 = 16{,}272{,}120{,}284$$

$$CV(\bar{T}) = \frac{\sqrt{16{,}272{,}120{,}284}}{1{,}006{,}831} = .1275 \text{ or } 12.75\% \ .$$

The relative errors for each of the three methods of selection in order of magnitude are:

12.75% (for method 2)

9.02% (for method 1)

7.05% (for method 3)

The simple random sample is much less efficient than either the systematically selected replicates over paper strata or a sample stratified by paper strata.

4.3.4 Exercise 7:

Each student is to compute the expanded estimates for table 4 based on each segment for the grouped strata and derive the estimated population total for the country and its sampling error.

Table 3--Reported tareas of coffee by segment and subsample

Subsample or replicate No.(m) / Segment or paper strata No.(k)	Intense coffee and cacao stratum (II)			Other stratum (I, III, IV, V)	
	1	2	3	1	2
1 (X_{hij})	1010	886	1687	1933	369
2	352	65	432	1	0
3	1801	4205	4737	0	0
4	1248	336	2828	0	265
5	200	70	426	8	0
6	1117	1071	2132	1352	23
7	449	153	729	0	427
8	730	75	1718	0	0
9	1145	650	781	0	0
10	508	350	169	0	1787
11	1664	775	4		
12	86	2128	373		
13	1462	884	490		
14	1186	369	506		
15	2432	2328	1890		
16	202	324	311		
17	137	342	809		
18	2217	3056	1220		
19	0	2351	1500		
20	30	381	252		
Total for replicated (X_{hi})	17976	20799	22994	3294	2871
Expansion factor for replicate	$\frac{N_1}{n_1} = \frac{978}{20} = 48.9$			$\frac{N_2}{n_2} = \frac{14917}{10} = 1491.7$	
Average replicate total ($\bar{X}_h$)	20,589.67			3,082.5	
Average expanded total for population ($\bar{T}_h$)	1,006,834.86			4,598,165.25	

Table 4--Expanded tareas of coffee by segment and subsample for one stratum

Subsample or replicate No.(m) / Segment or paper strata No.(k)	Intense coffee and cacao stratum			Other stratum	
	1	2	3	1	2
1	49,389.0	43,325.4	82,494.3		
2	17,212.8	3,178.5	21,124.8	To	
3	88,068.9	205,624.5	231,639.3		
4	61,027.2	16,430.4	138,289.2		
5	9,780.0	3,423.0	20,831.4		
6	54,621.3	52,371.9	104,254.8	be	
7	21,956.1	7,481.7	35,648.1		
8	35,697.0	3,667.5	84,010.2		
9	55,990.5	31,785.0	38,190.9		
10	24,841.2	17,115.0	8,264.1	completed	
11	81,369.6	37,897.5	195.6		
12	4,205.4	104,059.2	18,239.7		
13	71,491.8	43,227.6	23,961.0	by	
14	57,995.4	18,044.1	24,743.4		
15	118,924.8	113,839.2	92,421.0		
16	9,877.8	15,843.6	15,207.9	student	
17	6,699.3	16,723.8	39,560.1		
18	108,411.3	149,438.4	59,658.0		
19	0	114,963.9	73,350.0		
20	1,467.0	18,630.9	12,322.8		
Population total by replicates (T_h)	879,026.4	1,017,071.1	1,124,406.6	4,913,659.8	4,282,670.7
Average population total ($\bar{T}_h$)		1,006,834.7		4,598,165.3	

Table 5--Means and variances of paper strata for Stratum II

Paper strata	Means of paper strata	Variances for paper strata
1	1,194.3	185,884
2	283.0	37,243
3	3,581.0	2,447,056
4	1,470.0	1,589,701
5	232.0	32,452
6	1,440.0	359,677
7	443.7	82,965
8	841.0	684,103
9	858.7	65,780
10	342.3	28,774
11	814.3	690,060
12	862.3	1,222,026
13	945.3	239,017
14	687.0	191,443
15	2,216.7	82,737
16	279.0	4,489
17	429.3	118,616
18	2,164.3	844,804
19	1,283.7	1,416,900
20	221.0	31,521
Total all strata	20,589.6	10,355,251
Average per paper stratum	1,029.5	517,762

4.4 Tunisian Acreage and Livestock Survey

4.4.1 Background

The 1974 Tunisian Acreage and Livestock Enumerative Survey (Base Line Study) had its start with the Livestock Project initiated by AID in Tunisia, which is aimed at increasing livestock production through an extension program. For their project to be successfully evaluated at its conclusion, AID felt it would be necessary to obtain additional information concerning the current agricultural situation in Tunisia. In particular, they needed a study of farmers having livestock to obtain certain sociological information on the Tunisian livestock producer and the characteristics of livestock operations in the five northern gouvernorats. As the study developed, certain other information was deemed necessary and thus the following purposes for the survey evolved:

1. Develop a profile of livestock producers determining such items as the level of education, size of household, amount of hired labor, sources of agricultural information, ability to use agricultural information, etc.
2. Determine land use, the extent of crops complementary to livestock activity, availability of forages and grains for feed, and the extent of crops competitive to livestock, such as vegetables, fruits, and food grains.
3. Determine livestock base information, including inventories and classification of cattle and sheep.
4. Examine current practices of livestock producers, including feeding and grazing practices, housing practices, etc.

In late 1973 the Statistical Reporting Service of the U.S. Department of Agriculture was contacted by AID and a tentative project developed. A team of technical consultants consisting of three SRS employees was assigned to the project in January 1974. The team's first work in Tunisia occurred in late January. The bulk of the Tunisian personnel involved in the project was supplied by the National Institute of Statistics. They provided a group of nine technicians to construct the frame and select the sample and approximately 45 people to serve as field enumerators and supervisors. They also provided overall leadership in questionnaire design and conduct of the survey. Other agencies in Tunisia provided input in the scope of the survey, questionnaire development and limited personnel in the actual survey work.

The pretest survey was conducted in early March and provided good results. The actual survey took place in May 1974 with training of field personnel occurring the last week in April.

Four land use strata were distinguished for the five northern gouvernorats of Tunisia. The definitions of the four strata were:

I. Cultivated - All land which has been worked or improved in some manner was included in this stratum. The only exceptions are those areas where vineyards or orchards were extensive enough to be placed in a separate stratum.

II. Noncultivated - The major component of this stratum was the national forests; it generally consists of either woodland or brush. Mountains which for the most part have little or no vegetation were included. The personal knowledge of the Tunisian staff contributed greatly in each specific decision concerning the inclusion of noninterested land in this stratum.

III. Urban - This stratum consists of intensively populated areas such as cities and villages. Each city or village must have an area of 1/4 square kilometer or more to be included. The topographic maps which were used for stratification were unsatisfactory for delimiting the cities and villages because of their scale and age. Contact photo prints (scaled 1/12,500 and 1/25,000) were ordered for each city or village and used in the stratification process. The latest photographs available for this purpose were taken in 1962. It was recognized that the cities and villages may have expanded since the photos were taken. However, these problems may be treated on an individual basis whenever they arise.

IV. Arboriculture - All vine and tree crops specifically designated on the topographic maps were included in the stratum. It was recognized that these may have changed; however, it was determined that circumstances did not permit extensive field work to update the maps. Vineyards and orchards which were known to be cleared were omitted, but no new plantings were included.

After the preceding land use strata definitions were developed, the actual stratification process began. This consisted of segregating the land area on the maps according to the strata definitions using available physical boundaries such as roads, trails, streams, field edges when necessary, etc. After all land area was classified into one of the four strata, each stratum was subdivided into small areas ranging from 1/2 to 30 square kilometers, depending on the stratum definition. Each of the smaller areas within strata, referred to as count units, was planimetered to determine its area. Sampling units were assigned to each of the count units by dividing the area of the count unit by the desired area of the sampling unit and rounding the result to the nearest whole number. Sampling unit

size was approximately 2 square kilometers for Strata I, II, and IV, and 1/16 square kilometer for Stratum III.

The count units were then numbered in a serpentine order, starting in the upper right-hand corner of each map page, and a list was prepared showing the number of sampling units for each count unit. From this list samples of count units were selected with probabilities proportional to the number of sampling units in each. The count units were then subdivided into the assigned number of sampling units. A sampling unit was then randomly selected from those in the selected count unit. The selected sampling unit was the segment to be used in the field enumeration. All segments within a stratum had the same probability of selection.

At this time enlarged photographs at 1/6,250 or 1/12,500 scale were ordered. Upon receipt of the enlarged photographs the boundaries of the segments were transferred from the maps to the photographs. In some instances, it was apparent that the segment was too heavily populated for efficient enumeration and therefore they were again subdivided into equal parts, a random part selected for enumeration and the probability of selection adjusted. Maps showing the segment location, the enlarged photograph for each segment, and plastic grids made to the scale of the photographs were then included in kits for use by the field enumerators.

The Sampling Frame Configuration and Sample Size for the Initial Survey

Strata	Expected S.U. size	Total number S.U. N_h	Sample allocation (No. S.U.) = n_h	Number replicated samples = m_h	Number of paper strata (i.e. cluster size) = k_h
I	2 km^2	8,998	270	18	15
II	2 km^2	3,826	36	4	9
III	1/16 km^2	2,000	35	7	5
IV	2 km^2	286	18	6	3
			359		

4.4.2 Calculating Estimates and Sampling Errors

A summary of the expanded hectares of barley by segment and stratum are presented in table 6. The expanded hectares are derived by multiplying the reported segment data by 600 for Stratum I and 426 for Stratum II. The figures 600 and 426 represent the total number of frame units in the paper strata for the respective land use strata (i.e., primary strata). The estimate and its sampling error squared are derived using the same formulas as given for the Dominican Republic case.

Table 6--Expanded hectares of barley by segment, replicate and stratum

Segment or paper strata No. (k)	Stratum I			Stratum II	
	Replicate (m)			Replicate (m)	
	1	2	3	1	2
1 ($\hat{X}_{hij}$)	633.6	4,346.1	1,207.8	932.5	8,206.0
2	0	782.1	0	596.8	0
3	3,148.2	0	821.7	2,945.7	0
4	396.0	891.0	514.8	8,392.5	857.9
5	0	0	0	0	0
6	158.4	1,247.4	1,386.0	0	0
7	99.0	0	0	2,890.8	186.5
8	148.5	277.2	0	2,704.3	167.8
9	574.2	1,326.6	722.7	1,678.5	2,424.5
10	0	0	0		
11	198.0	0	990.0		
12	1,178.1	643.5	475.2		
13	693.0	0	574.2		
14	1,881.0	2,871.0	613.2		
15	1,009.8	831.6	1,435.5		
Population total by replicate ($\hat{T}_{hi}$)	10,117.8	13,731.3	8,741.7	20,142.1	11,842.7
Average population total ($\bar{\hat{T}}_h$)		10,863.6		15,992	
$V(\bar{\hat{T}}_h)$		2,213,730.1		34,440,020.2	
$CV(\bar{\hat{T}}_h)$		13.6%		36.6%	

4.4.3 Postsurvey Analysis

As was done for the Dominican Republic, alternative methods of selecting samples from the frame are considered. In Stratum I there are 15 paper strata, each consisting of 600 sampling units; i.e., 8998 ÷ 15 = 599.9. Three segments were selected from each paper stratum. Each sampling unit or segment provides an estimate of the total for the paper strata, and each replicate (cluster of 15 sample units) provides an estimate of the population total.

Method 1: A variance is computed for each paper stratum from the m segments and these variances are added:

$$V(T_h) = \sum_{j=1}^{15} V(T_{hj})$$

$$V(T_h) = 9,675,283.1$$

$$V(\bar{T}_h) = \frac{9,675,283.1}{3} = 3,225,094.4$$

and the average variance within paper strata for expanded data is:

$$V(T_{hw}) = \sum_{j=1}^{15} W_k V(T_{hj}) = \frac{9,675,283.1}{15} = 645,018.9$$. (See last line, col. 3, table 5.)

Method 2: The postsurvey analysis of expected sampling error is based on selecting a single simple random sample of size 45:

$$V(T) = \sum_{j=1}^{15} W_k S_k^2 - \sum_{j=1}^{15} \frac{W_k S_k^2}{m_k} + \sum_{j=1}^{15} \frac{W_k^2 S_k^2}{m_k} + \sum_{j=1}^{15} W_k \bar{x}_k^2 - \left(\sum_{j=1}^{15} W_k \bar{x}_k \right)^2$$

(Cochran 2nd Ed. p. 139)

$$= \frac{9,664,257}{15} - \frac{9,664,257}{15(3)} + \frac{9,664,257}{15^2(3)} + \frac{13,428,742}{15} - \frac{13,428,742}{15^2}$$

= 1,279,406 (the within-stratum variance based on expanded segment data and assuming a simple random sample had been selected)

$$V(x_i) = \frac{V(T)}{N_i^2} = \frac{1,279,405}{600^2} = 3.553906$$ (the variance for a sample of size 1 if unexpanded data had been used in calculating paper strata variances and means)

or

$$V(\bar{x}) = \frac{3.553906}{45} = .0789757 \text{ is the variance for a sample of size 45}$$

and

$$V(\bar{T}) = N^2 V(\bar{x}) = (8,998)^2\ (.0789757) = 6,394,188$$ (variance of mean expanded total assuming simple random sample)

Method 3:

$$V(T) = \frac{\sum_{i=1}^{3} T_i^2 - \frac{(\sum T_i)^2}{3}}{3 - 1}$$

$$= \frac{(10,117.8)^2 + (13,731.3)^2 + (8,741.7)^2 - \frac{(32,590.8)^2}{3}}{3 - 1}$$

$$= \frac{13,282,380.59}{2} = 6,641,190.15 \text{ for 1 replicate.}$$

$$V(\bar{T}) = \frac{6,641,190.27}{3} = 2,213,730.05 \text{ for mean of 3 replicates.}$$

The three variances we wish to compare based on the three different proposed methods of sampling are as follows:

Method 1:

$V(\bar{T}_h) = 3,225,094$ (selecting a simple random sample of size 3 in each of the 15 paper strata within Stratum I; i.e., n = 3 x 15 = 45)

Method 2:

$V(\bar{T}_h) = 6,394,188$ (selecting a simple random sample of size 45 out of stratum I; i.e., n = 45)

Method 3:

$V(\bar{T}_h) = 2,213,730$ (systematically selecting three replicates of size 15 based on paper strata in Stratum I; i.e., n = 15 x 3 = 45)

These results are similar to those for the Dominican Republic. The coefficient of variation with simple random sampling is approximately twice as

large as that of stratified systematic sampling with paper strata; that is, 23.2% to 13.3%. In both cases presented in this chapter, the use of equal-size paper strata substantially reduced the variance of the estimates as compared with simple random sampling without paper strata.

4.4.4 Exercise 8:

Using the results in table 6, each student is to compute the estimated population total and its sampling error.

CHAPTER 5 - USE OF SEVERAL FRAMES IN SAMPLING

5.1 Introduction

In this chapter, we introduce a general methodology for "multiple-frame surveys." The need for several frames arises because: (1) the individual frames do not completely cover all the units in the population but collectively the frames do include all the population units of interest, or (2) even though all the units in the population of interest are covered by a single frame, the use of several frames leads to smaller expected sampling errors per dollar spent. In either case, the use of several frames results in some units being included in more than one frame. For the population units common to several frames, two or more estimators of the same parameter are available. The material covered in this chapter deals with the general theory of utilizing two frames with and without prior knowledge as to the extent of their mutual overlap. The technique of domain estimation is employed. The "overlap domain" provides two estimates of the same parameter, one from each frame. Consequently, it is necessary to test the reasonableness of the assumption that the two sample estimates of the parameter (i.e., means) are equal; that is, the hypothesis that the estimates of the parameter value are equal must be tested before the estimates can be "pooled." In the event the assumption of equality of the parameter is rejected, the sample data do not suggest which frame should be used to obtain the estimate of the parameter but do indicate that any two-frame estimate for the total population does not satisfy all the basic assumptions. The decision of which frame to base the estimates on must be based on other statistical considerations.

Aside from the theoretical considerations of sampling, multiple-frame surveys are more difficult to execute operationally and require more controls to prevent nonsampling errors becoming large. This is a direct result, in practice, of the survey procedures and structure of each frame (which may consist of different types of listing or frame units); in addition, the elementary reporting units (i.e., individuals, farm operators or land tracts) themselves may differ from one frame to another. Thus, operationally the survey may include two frames with different types of listing or frame units, two different types of elementary units, two different procedures for associating the population of interest with the sampling units, and of identifying all units in the population of interest which are in both frames.

5.2 Example of Representing a Population in Two Frames

A small population of farms is used to illustrate the relationship between single-frame surveys and the two-frame surveys. A farm weight is shown explicitly for the individual farms in each frame as well as for both frames since there may be a possiblility of duplication of farm units within a frame as well as between frames. The recognition of duplication of farms is critical in deriving unbiased estimators for both single and two-frame surveys. In all cases, the individual farm weight must be equal to 1.0. This is clear from the fact that, if either a single frame or both frames are completely enumerated, each farm must be included just once and only once, if unbiased estimates are to be obtained. First, the population of interest, i.e., "farms," must be defined. One definition is as follows: A farm is defined as a unit of land managed by one or more individuals to produce one or more agricultural products. If several individuals (i.e., more than one) are involved in managing a farm unit, only the jointly managed land producing agricultural products is part of a particular farm unit. Other land managed by either one of the individuals or part of the management group to produce agricultural products will be considered as different or separate farm units. That is, a unique management group exists for each farm unit in the population of interest. Finally, those crops or livestock species which will be considered as agricultural products need to be defined. However, for our purposes these agricultural products need not be specifically enumerated but are understood to consist of products intended to be consumed as food by people or of products of the land used to produce food consumed by people. It should be clear that the population of interest which is defined will have a direct bearing on the multiple frame procedures specified.

The population of farms as related to the frame units in the Frame A (i.e., area frame) is shown on page 103. The frame units are area segments and each farm is associated with one and only one frame unit by a unique headquarters rule based on the operator's residence and there is no duplication of farms within the frame. If there is more than one person involved in the management group, only the "senior" person (or officer) will be considered the operator. Frame A contains all the farms in the population of interest.

The second frame, B, is a name-address list given on page 104 which gives access to an unknown portion of the population of farms since the name-address units may represent individuals, corporate names, partnership names, persons

no longer farming, or may represent individuals who never operated a farm but were erroneously included in the list. A farm which is jointly operated (i.e., a management group of two or more) may be duplicated through more than one frame unit if a farm association rule is employed which permits more than one member of a management group to be an operator of the farm unit through different name-address units. That is, a farm unit can be accessed by different name-address units in Frame B, but only one farm unit is associated with each frame unit. When a farm unit may be associated with several frame units that have different names and addresses then provision must be made to detect the number of units duplicated in this way by means of the survey questionnaire. The following is an example of a name-address list procedure in which both the name and address are required to define the frame unit.

Frame Unit	Farm Associated
(1) A Personal Name - Residential Address	A farm unit operated only by individual listed at address
(2) A Personal Name - Business Address	A farm unit operated only by individual listed at address
(3) "Corporate" Name - Business Address	"Corporate" farm unit only, and no other farm units of persons at same name-address unit
(4) "Corporate" Name - Residential Address	"Corporate" farm unit only, and no other farm units of persons at same name-address unit

The word "corporate" can be replaced by the words partnership, estate, cooperative or institution to cover additional types of situations which may occur. Where the name portion of the unit consists of two or more personal names, these frame units will be treated as partnerships. A business address is any address which does not correspond to a residence.

The illustration below uses the procedures for associating farm units with frame units based on one set of rules for each frame for a given definition of a farm (and also for not associating farm units with frame units). For the duplication within a frame, weights (w_j) are assigned to the frame units for the i^{th} farm so they add to 1.0 for each farm in the population of interest. In two-frame surveys the idea of duplication between frames is handled by partitioning the farms into three domains: (a) farms in Frame A only, (b) farms in Frame B only, and (ab) farms in both frames A and B. In this illustration it is assumed that there are no farms in Domain (b). In the development of the theory by Hartley and others, it has been assumed that N_A and N_B, the number of units in the population of interest for each frame, are known. In addition, the number

of units for the population of interest in each domain, N_a, N_b, and N_{ab}, may be assumed to be known. The between-frame duplication is handled by assigning frame weights, p and q for Domain (ab) where $p, q \geq 0$ and $p + q = 1$, and no frame weights for Domain (a) or (b) because there is no duplication between frames.

The combined result of the farm weights ($\sum_i w_{ij.}$) for the single frames, A and B, and the frame weights (p and q) for Domain (ab) is such that $p\sum_i w_{jA} + q\sum_i w_{jB} = 1$ for each farm, and for Domain (a), $\sum_i w_{jA} = 1.0$ and for (b) $\sum_i w_{jB} = 1.0$ where the subscript i refers to a farm while the subscripts A and B refer to the individual frames and the summation is for the i^{th} farm over the j frame units which the i^{th} farm is associated with in each frame.

The farm weights $\sum_i w_{jA}$ and $\sum_i w_{jB}$ are the result of the procedure for associating farms with the frame units and are generally determined from information obtained on the questionnaire with appropriate checking of questionable name-address units and follow-up inquiry to the farm. The frame weights, p and q, can be derived based on several criteria as follows:

(1) Optimum weights which have minimum variance per dollar spent for each survey characteristic;

(2) A single set of weights for all survey characteristics to minimize the computational problems and to insure that subgroups will add to the group total;

(3) Frame weights derived independently of the survey characteristics being estimated to assure unbiased estimates;

(4) Frame weights constant between successive surveys for the same characteristics so that the change in the estimates over time is not due to different weighting schemes (i.e., assuming the same frames are used).

The derivation of optimum frame weights will be discussed later, but several simple schemes which satisfy criteria (2), (3) and (4) above are given even though the efficiencies of the schemes are not known except for special cases.

(1) p and q based on constant sampling fractions:
The sampling fraction is f_A for Frame A and f_B for Frame B. The frame weights would be

$$p = \frac{f_A}{f_A + f_B}$$ where p is the weight for farms in Domain (ab)

which are accessed by Frame A, and

$$q = \frac{f_B}{f_A + f_B}$$ where q is the weight for farms in Domain (ab)

which are accessed by Frame B.

(2) p and q based on frame sizes:

The number of frame units is F_A for Frame A and F_B for Frame B. The corresponding frame weights would be

$$p = \frac{F_A}{F_A + F_B}, \quad \text{and}$$

$$q = \frac{F_B}{F_A + F_B}.$$

(3) p and q based on constant sample sizes:

The sample sizes are n_A for Frame A and n_B for Frame B. The corresponding weights would be

$$p = \frac{n_A}{n_A + n_B}, \quad \text{and}$$

$$q = \frac{n_B}{n_A + n_B}.$$

(4) p and q based on sample sizes, n_A and n_B, and cost of data collection per frame unit, C_A and C_B.

$$p = \frac{\frac{n_A}{C_A}}{\frac{n_A}{C_A} + \frac{n_B}{C_B}}$$

$$q = 1 - p$$

5.2.1 Population Units Related to Frame A

The following table shows the population of 10 farms accessed by Frame A and the relationship between the frame units, the farms, and the farm weights.

Farm No.	Frame Unit No. (i.e. segment)	Units Accessible from Frame		
		Value of Farm Characteristic (Hectares of Wheat)	Individual Farm Weight for Frame Unit	Tabulated Value of Farm Characteristic for Frame Unit (Value x Weight)
1	1	10	1.0	10
2	2	7	1.0	7
3	2	11	1.0	11
4	3	20	1.0	20
5	3	3	1.0	3
6	4	0	1.0	0
7	5	9	1.0	9
8	6	26	1.0	26
9	7	0	1.0	0
10	7	14	1.0	14
Frame Total	7		10	100

5.2.2 Population Units Related to Frame B

The portion of the population of farms accessed by the Name-Address Frame and the relationship between frame units, the farms, and the farm weights is given in the following table.

Farm No.	Frame Unit No. (Each Different Name-Address)	Units Accessible from Frame: Value of Farm Characteristic (Hectares of Wheat)	Individual Farm Weight for Frame Unit	Tabulated Value of Farm Characteristic for Frame Unit (Value x Weight)
3	1	11	.5 1/	5.5
3	2	11	.5 1/	5.5
			(1.0)	
4	3	20	1.0	20
7	4	9	1.0	9
8	5	26	1.0	26
9	6	0	1.0	0
No Farm	7 2/	0	0	0
No Farm	8 2/	0	0	0
Frame Total	8		5	66

1/ Farm No. 3 is a partnership. The name-addresses of the two individual partners were included in the list because the list sources were not aware of the partnership. The survey questionnaire would have to detect the partnership if either one of the frame units was selected in a sample.

2/ There is no farm associated with this frame unit because the name-address listing does not qualify as a farm based on either the rules of associating farms with frame units or the definition of a farm. However, the list source was not aware of this fact; consequently, a survey procedure must be used to determine this for each frame unit sampled.

5.2.3 Population Units Related to Frames A and B

The population of 10 farms is accessed by Frames A and B. The relationship between frame units, farm weights to remove duplication and value of survey characteristics for a two-frame survey are shown in the following table. In this illustration the values of p and q are derived from the frame sizes $F_A = 7$ and $F_B = 8$. That is, $p = \frac{7}{7 + 8} = .4667$ and $q = 1 - p = .5333$.

Farm No.	Farms Accessible from Area Frame						Farms Accessible from List Frame				Farm Weight for Two Frames
	Domain (a)			Domain (ab)			Domain (ab)				
	Frame Unit No.	Hectares of Wheat	Individual Farm Wt., (i.e., No. Farms)	Tabulated Value for Wheat 1/	Hectares of Wheat	Individual Farm Wt. x p, (i.e. No. Farms)	Tabulated Value for Wheat 1/	Frame Unit No.	Hectares of Wheat	Individual Farm Wt. x q, (i.e. No. Farms)	Tabulated Value for Wheat 2/	
1	1	10	1.0	10								1.0
2	2	7	1.0	7								1.0
3	2				11	.4667	5.1337	1 2	11 11	.26665 .26665	2.93315 2.93315	1.0
4	3				20	.4667	9.3340	3	20	.5333	10.6660	1.0
5	3	3	1.0	3								1.0
6	4	0	1.0	0								1.0
7	5				9	.4667	4.2003	4	9	.5333	4.7997	1.0
8	6				26	.4667	12.1342	5	26	.5333	13.8658	1.0
9	7				0	.4667	0	6	0	.5333	0	1.0
10	7	14	1.0	14								
-	1				0	0	0	7	0	0	0	0
-	5				0	0	0	8	0	0	0	0
Frame Total	7		5	34		2.3335	30.8022	8		2.6665	35.1978	10

1/ Value x farm weight x p, 2/ Value x farm weight x q.

The two-frame estimates for number of farms and hectares of wheat, if each frame is completely enumerated, are as follows by frames:

Frame and Domain	Survey Characteristics	
	Number of Farms	Hectares of Wheat
Frame A - Domain 1 (a)	5	34
Frame A - Domain 3 (ab)	2.3335	30.8022
Frame B - Domain 3 (ab)	2.6665	35.1978
Frame B - Domain 2 (b)	0	0
Frame A and B	10.0000	100.0000

The two-frame estimator for the total of a survey characteristic can be written in equation form:

$$(1)\quad \hat{Y} = F_A \bar{y}_1 + F_A\, p\, \bar{y}_{3A} + F_B\, q\, \bar{y}_{3B} + F_B \bar{y}_{2B}$$
$$= 7\, \bar{y}_1 + 7(.4667)\, \bar{y}_{3A} + 8(.5333)\, \bar{y}_{3B} + 0 \quad ,$$

where F_A, F_B, p, q were defined previously and

$\bar{y}_1$ = mean per frame unit in Domain (a) in Frame A

$\bar{y}_{3A}$ = mean per frame unit in Domain (ab) in Frame A

$\bar{y}_{3B}$ = mean per frame unit in Domain (ab) in Frame B

$\bar{y}_{2B}$ = mean per frame unit in Domain (b) in Frame B.

It should be noted that in this estimator the number of farms in Frame A, N_A, and the number of farms in Frame B, N_B, are not assumed known and are not used in the estimator. We have assumed that only the number of frame units are known and these are sampled. The two-frame estimator for the total of a survey characteristic could also be written as follows:

(2) If N_A and N_B were known

$$\hat{Y} = N_A\, \bar{y}'_1 + N_A\, p\, \bar{y}'_{3A} + N_B\, q\, \bar{y}'_{3B} + N_B\, \bar{y}'_{2B} \quad , \text{ or}$$

(3) If N_a, N_{ab}, and N_b were known

$$\hat{Y} = N_a \bar{y}'_1 + N_{ab}\, p\, \bar{y}'_{3A} + N_{ab}\, q\, \bar{y}'_{3B} + N_b \bar{y}'_{2B}$$

where $\bar{y}'_1$ = mean per farm in Domain (a) in Frame A

$\bar{y}'_{3A}$ = mean per farm in Domain (ab) in Frame A

$\bar{y}'_{3B}$ = mean per farm in Domain (ab) in Frame B

$\bar{y}'_{2B}$ = mean per farm in Domain (b) in Frame B

In the example, $N_A = 10$ and $N_B = 5$ while $N_a = N_{ab} = 5$, and $N_b = 0$.

Table 1 below reveals several interesting and unusual results which arise because of properties of the two frames and the survey characteristics. The range in the estimates of total number of farms and hectares of wheat is somewhat greater for the two-frame survey. This is largely the result of the list frame not being very efficient for this population of farms when the total number of farms is not known in advance. The total number of frame units sampled is 2 (i.e., n = 1 for each frame) for the two-frame survey as compared to only 1 for the survey from Frame A.

Table 1--Range of Estimates for Population of 10 Farms Based on Samples of Size 1 from Frame A and Frame B Using Estimator (1)

Survey Characteristic	Frame A Survey	Two-Frame Survey
Total Number of Farms		
Mean Total	10	10
Range of Estimates	14 - 7 = 7	14.5333 - 3.2669 = 11.2664
Total Hectares of Wheat		
Mean Total	100	100
Range of Estimates	182 - 0 = 182	195.8658 - 0 = 195.8658
Number of Estimates	7	56
Number of Frame Units	1	2

5.2.4 Variance of Two-Frame Estimator

The variance of the estimated total for a characteristic of the population of farms is given for the first estimator. Independent random samples of the frame units, n_A and n_B, are selected from each frame. The sampling of frame units results in a random sample of farm units being selected from each frame but the sample size in terms of number of farm units is unknown at the time of selection. The first estimator in Section 5.2.3 was based on the number of frame units, F_A and F_B, being known and the number of farm units, N_A and N_B, being unknown. The variance of this two-frame estimator is:

$$(1)\quad V(\hat{Y}) = F_A^2 \frac{V(y_1)}{n_A} + F_A^2 p^2 \frac{V(y_{3A})}{n_A} + F_A(F_A-1)p^2 \frac{Cov(y_1, y_{3A})}{n_A}$$

$$+ F_B^2 p^2 \frac{V(y_{3B})}{n_B} + F_B^2 \frac{V(y_2)}{n_B}$$

where the variances of y_1, y_{3A}, y_{3B}, y_2 are computed as the variance of a simple random sample using the domain totals for each selected frame unit.

The covariance term arises in the variance estimator because the frame units in Frame A are clusters of farm units (i.e. segments). A segment in Frame A may contain farms in both domains (a) and (ab). Consequently, a correlation or nonzero covariance term may exist between farms in domains (a) and (ab) in Frame A. However, the expected value of the statistical correlation between the sample farms in Frames A and B is zero if independent random samples are selected from each frame.

5.2.5 Exercises 9 and 10

Exercise 9:

Each student should define a different procedure (or rule) for associating farms with the units in each frame.

Exercise 10:

Each student should define a new agriculture unit for the population of interest.

5.3 Two-Frame Theory

The technique employed is that of domain estimation which has been set forth by Hartley and others. The theory has been developed assuming either the number of units in the population of interest is known for both frames or for each domain. In the example in the preceding section, the total number of farms in each frame was assumed to be unknown, but could have been estimated from the sample.

5.3.1 Two-Frame Methodology

Consider two frames, A and B, and assume that a sample has been drawn from each frame. The samples may be entirely different in the two frames but the following assumptions are made:

(1) Every unit in the population of interest belongs to at least one of the frames.

(2) It is possible to record for each sampled unit in each frame whether or not it belongs to the other frame.

This means we can divide the units of the sample into three $(2^2 - 1)$ domains.

Domain (a) The unit belongs to Frame A only.

Domain (b) The unit belongs to Frame B only.

Domain (ab) The unit belongs to both frames.

The units in the population are also conceptually divided into the above domains.

5.3.2 Notation for Two-Frame Surveys

There are four different situations concerning our state of knowledge of the total number of units in the frame, population, and in the domains as well as our ability to allocate prescribed sample sizes to the domains. We consider only cases 1, 2, and 3 in the discussion. In case 4, the sample sizes in terms of the population of interest are random variables since the number of population units in each frame is unknown. This case was discussed in Section 5.2. Unless otherwise stated, the type of elementary unit is the same in both frames.

Table 2--Notation

	Frame		Domain		
	A	B	a	b	ab
Number Frame Units	F_A	F_B			
Population Number	N_A	N_B	N_a	N_b	N_{ab}
Sample Size (Population Units)	n_A	n_B	n_a	n_b	n_{ab} & n_{ba}
Population Total	Y_A	Y_B	Y_a	Y_b	Y_{ab}
Population Mean	$\bar{Y}_A$	$\bar{Y}_B$	$\bar{Y}_a$	$\bar{Y}_b$	$\bar{Y}_{ab}$
Sample Total	y_A	y_B	y_a	y_b	y_{ab} & y_{ba}
Sample Mean	$\bar{y}_A$	$\bar{y}_B$	$\bar{y}_a$	$\bar{y}_b$	$\bar{y}_{ab}$ & $\bar{y}_{ba}$
Cost of Sampling Unit	C_A	C_B			

Random samples are drawn from each frame and n_{ab} and n_{ba} are the subsamples of n_A and n_B, respectively, which fall into the overlap Domain (ab) where the first letter a or b indicates the frame from which the sample was drawn. The means $\bar{y}_{ab}$ and $\bar{y}_{ba}$ can be computed only if $n_{ab} > 0$ and $n_{ba} > 0$.

Table 3--Four Cases of Prior Knowledge

Case	Knowledge of Population Numbers in Domains and Frames	Possibility of Fixed Sample Allocations to Domains & Frames	Nature of Domains
1	$N_A, N_B, N_a, N_b, N_{ab}, F_A, F_B$ Known	It is feasible to allocate sample sizes to domains	Domains = Strata
2	$N_A, N_B, N_a, N_b, N_{ab}, F_A, F_B$ Known	It is not feasible to allocate sample sizes to domains	Domains = post-strata
3	Only N_A, N_B, F_A, F_B Known	Sample sizes can only be allocated to frames	Domains = domains proper
4	$N_A, N_B, N_a, N_b, N_{ab}$ Unknown F_A, F_B Known	Sampling rates only can be allocated to frames	Domains = domains in populations of unknown sizes

5.3.3 Estimation of Population Totals and Means

In case 1 the estimation problem is reduced to the standard methodology for stratified sampling. For cases 2 and 3 two approaches leading to identical formulas are possible: (a) the theory of domain estimation, or (b) the method of weighted variables. For (b) we introduce the following attributes to units in the two frames for a survey item denoted by y_i:

$$\text{Frame A} \quad y_i' = \begin{cases} y_i \text{ if } i^{th} \text{ unit is in Domain (a)} \\ c_i y_i \text{ if } i^{th} \text{ unit is in Domain (ab)} \end{cases}$$

$$\text{Frame B} \quad y_i' = \begin{cases} y_i \text{ if } i^{th} \text{ unit is in Domain (b)} \\ d_i y_i \text{ if } i^{th} \text{ unit is in Domain (ab)} \end{cases}$$

where c_i and d_i are numbers which satisfy for each population unit in Domain (ab) $E(c_i + d_i) = 1$. Therefore, the two frames are to be converted into two mutually exclusive strata of sizes N_a and N_{ab} for Frame A and N_b and N_{ab} for Frame B. That is, we have duplicated the N_{ab} units in both frames. The population total will be equivalent to the single frame total of Y. However, the sample estimator of the total and the variance are easily derived only if c_i and d_i are constants. That is, $c_i = p$ and $d_i = q$ where $p + q = 1$ and are determined independently of the parameter being estimated for unbiasedness. Clearly, the population total is equivalent to the original population since the $N = N_a + N_{ab} + N_b$ units are now $N_a + (p + q) N_{ab} + N_b$ and the totals are:

$$Y = Y_a + Y_{ab} + Y_b \text{ , or}$$

$Y' = Y_a + pY'_{ab} + qY'_{ba} + Y_b$ where there are two independent estimators of Y_{ab} which are combined.

The standard methodology applicable to the survey designs in Frame A and Frame B is used to obtain estimates of the two stratum totals for the variate y_i', their variances and variance estimates. Adding the totals for both frames, we obtain the total for the population of interest. To obtain estimates of the population mean, $\bar{Y} = Y/N$, it is necessary to apply these same formulas to a new variable, called the count variable,

μ_i to estimate the population size, N, where the variable μ_i has either the value 0 or 1.

The estimate of the population total given by Hartley for a characteristic when N_a, N_b and N_{ab} are known is:

$$\hat{Y} = N_a\bar{y}_a + N_{ab}p\bar{y}_{ab} + N_{ab}q\bar{y}_{ba} + N_b\bar{y}_b \quad .$$

This estimator is in the form of a post-stratified sampling estimator. If the sample is sufficiently large and the f.p.c. factor is not important, the variance is given by

$$V(\hat{Y}) = \frac{N_A^2}{n_A}\left[\sigma_a^2 N_a + \sigma_{ab}^2 N_{ab} p^2\right] + \frac{N_B^2}{n_B}\left[\sigma_b^2 N_b + \sigma_{ab}^2 N_{ab} q^2\right]$$

where σ_a^2, σ_b^2 and σ_{ab}^2 are the within post-stratum variances.

When N_a, N_b, and N_{ab} are unknown but N_A and N_B are known, an estimator given by Lund based on the actual subdivisions n_{ab} and n_{ba} is:

$$\hat{Y} = \frac{N_A}{n_a} n_a\bar{y}_a + \left[\frac{N_A}{n_A} n_{ab} p + \frac{N_B}{n_B} n_{ba} q\right]\bar{y}_{ab} + \frac{N_B}{n_B} n_b\bar{y}_b$$

where $\bar{y}_{ab} = \dfrac{n_{ab}\bar{y}_{ab} + n_{ba}\bar{y}_{ba}}{n_{ab} + n_{ba}}$. The approximate variance where $\alpha = N_{ab}/N_A$ and $\beta = N_{ab}/N_B$ is:

$$V(\hat{Y}) = \frac{N_A^2}{n_A}(1-\alpha)\sigma_a^2 + \frac{N_A N_B}{\alpha n_A + \beta n_B}\sigma_{ab}^2 + \frac{N_B^2}{n_B}(1-\beta)\sigma_b^2$$

$$+ \frac{N_A^2(1-\alpha)\alpha}{n_A}\left[\bar{Y}_a - p\bar{Y}_{ab}\right]^2 + \frac{N_B^2(1-\beta)\beta}{n_B}\left[\bar{Y}_b - q\bar{Y}_{ab}\right]^2$$

An alternative approach proposed by Fuller and Burmeister uses a multiple regression type estimator for samples selected from two overlapping frames. It is assumed that the sampling is such that unbiased estimators of the item totals and the total number of units in each domain are available. The estimator suggested for the population total of the content item is as follows:

$$\hat{Y} = \hat{Y}_a + \hat{Y}_B + \hat{\beta}_1(\hat{N}_{ab} - \hat{N}_{ba}) + \hat{\beta}_2(\hat{Y}_{ab} - \hat{Y}_{ba}) \text{ where } \hat{Y}_B = \hat{Y}_b + \hat{Y}_{ba} .$$

When Frame B is complete and Frame A incomplete, we do not have Domain (a), hence the estimator is

$$\hat{Y} = \hat{Y}_B + \hat{\beta}_1(\hat{N}_{ab} - \hat{N}_{ba}) + \hat{\beta}_2(\hat{Y}_{ab} - \hat{Y}_{ba})$$

where

$\hat{Y}_B$ = an unbiased estimator of the total constructed from the sample in Frame B,

$\hat{Y}_{ab}$ = an unbiased estimator of the total of Domain (ab) constructed from the sample of Frame A,

$\hat{Y}_{ba}$ = an unbiased estimator of the total of Domain (ab) constructed from the sample of Frame B,

$\hat{N}_{ab}$ = an unbiased estimator of the number of observational units in Domain (ab) constructed from the sample of Frame A,

$\hat{N}_{ba}$ = an unbiased estimator of the number of observational units in Domain (ab) constructed from the sample of Frame B, and

$\hat{N}_b$ = an unbiased estimator of the number of observational units in Domain (b) constructed from Frame B.

The optimal values of $\hat{\beta}_1$ and $\hat{\beta}_2$ are given by

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{bmatrix} \hat{V}(\hat{N}_{ab} - \hat{N}_{ba}) & \hat{Cov}(\hat{N}_{ab} - \hat{N}_{ba}, \hat{Y}_{ab} - \hat{Y}_{ba}) \\ \hat{Cov}(\hat{N}_{ab} - \hat{N}_{ba}, \hat{Y}_{ab} - \hat{Y}_{ba}) & \hat{V}(\hat{Y}_{ab} - \hat{Y}_{ba}) \end{bmatrix}^{-1} \begin{bmatrix} -\hat{Cov}(\hat{Y}_a, \hat{N}_{ab}) + \hat{Cov}(\hat{Y}_b, \hat{N}_{ba}) \\ -\hat{Cov}(\hat{Y}_a, \hat{Y}_{ab}) + \hat{Cov}(\hat{Y}_b, \hat{Y}_{ba}) \end{bmatrix}$$

A consistent estimator of the variance is

$$\hat{V}(\hat{Y}) = \hat{V}(\hat{Y}_a) + \hat{V}(\hat{Y}_b) + \hat{\beta}_1[\hat{Cov}(\hat{Y}_a,\hat{N}_{ab}) - \hat{Cov}(\hat{Y}_b,\hat{N}_{ba})]$$

$$+ \hat{\beta}_2[\hat{Cov}(\hat{Y}_a,\hat{Y}_{ab}) - \hat{Cov}(\hat{Y}_b,\hat{Y}_{ba})] .$$

It is also suggested that if other y characteristics are observed in the survey, it may be possible to further decrease the variance of the estimator by including other unbiased estimators of zero [1/] in the regression type equation.

5.3.4 Determination of Fixed Weights (p and q) for One-Survey Characteristic

The value of p is to be determined independently of the parameter being estimated, $\bar{Y}$ or Y. Assuming a simple cost function $C = C_A n_A + C_B n_B$ where C is the total cost of sampling, C_A is the cost of an observation from Frame A and C_B is the cost of an observation from Frame B. After some labor, the optimum value of p was found by Hartley to be one of the solutions of:

$$p^2\rho[\phi_B(1 - \beta) + \beta q^2] = q^2 [\phi_A(1 - \alpha) + \alpha p^2]$$

where

$$\rho = \frac{C_A}{C_B} , \quad \phi_B = \frac{\sigma_b^2}{\sigma_{ab}^2} , \quad \phi_A = \frac{\sigma_a^2}{\sigma_{ab}^2} , \quad \alpha = \frac{N_{ab}}{N_A} \quad \text{and} \quad \beta = \frac{N_{ab}}{N_B} .$$

Once the value of p has been determined, the values of n_A and n_B can be found from

$$\frac{n_A}{N_A} = \theta[(\sigma_a^2(1 - \alpha) + \alpha p^2\sigma_{ab}^2)/C_A]^{\frac{1}{2}}$$

$$\frac{n_B}{N_B} = \theta[(\sigma_b^2(1 - \beta) + \beta q^2\sigma_{ab}^2)/C_B]^{\frac{1}{2}}$$

1/ Refers to estimation theory.

where 0 would be determined by the budget available. The foregoing derivation requires knowledge of the costs, variances, and population domain sizes N_a, N_b, and N_{ab}.

An alternative derivation for p (due to Lund), when N_{ab} is known, is given by the simpler solution for p by the expression

$$p = \frac{\alpha n_A}{\alpha n_A + \beta n_B} .$$ The ratio, $\frac{n_A}{n_B}$, can be expressed by the

iterative system

$$r_1 = \sqrt{\frac{C_B}{C_A}} \left(\frac{\beta}{\alpha}\right) \quad \text{and} \quad r_{i+1}^2 = \frac{C_B}{C_A} \left(\frac{\beta}{\alpha}\right)^2 \frac{(r_i + \frac{\beta}{\alpha})^2 (1 - \alpha)\sigma_a^2 + r_i^2 \sigma_{ab}^2}{(r_i + \frac{\beta}{\alpha})^2 (1 - \beta)\sigma_b^2 + (\frac{\beta}{\alpha})^2 \beta \sigma_{ab}^2}$$

where $r_i \to r$. Limited experience with this method indicates it converges rapidly to a value for p (i.e., 2 to 5 iterations). Thus, the optimum value for p is the ratio of the expected value of the "overlap domain" size in Frame A with respect to the sum of the expected values of the "overlap domain" in both frames.

When N_a, N_b and N_{ab} are unknown, it is necessary to insert unbiased estimates of these three parameters. The minimization of the variance expression at the bottom of page 112 as a function of p, n_A and n_B subject to the cost equation specifies the optimum value (p_0):

$$p_0 = \frac{\frac{N_A(1-\alpha)}{n_A} \bar{Y}_a + \frac{N_B(1-\beta)}{n_B} (\bar{Y}_{ab} - \bar{Y}_b)}{\left[\frac{N_A(1-\alpha)}{n_A} + \frac{N_B(1-\beta)}{n_B}\right] \bar{Y}_{ab}}$$

The sample allocation among the two frames can be expressed by an **iterative** system

$$r_1 = \sqrt{\frac{C_B}{C_A}} \cdot \left(\frac{\beta}{\alpha}\right)$$

$$r_{i+1}^2 = \frac{C_B}{C_A}\left(\frac{\beta}{\alpha}\right)^2 \left\{ \frac{(1-\alpha)\sigma_a^2 + \frac{r_i^2\alpha\sigma_{ab}^2}{(r_i+\frac{\beta}{\alpha})^2} + \frac{r_i^2\alpha(1-\alpha)(\bar{Y}_a+\bar{Y}_b-\bar{Y}_{ab})^2}{[r_i+\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2}}{(1-\beta)\sigma_b^2 + \frac{(\frac{\beta}{\alpha})^2\beta\sigma_{ab}^2}{(r_i+(\frac{\beta}{\alpha})^2} + \frac{[\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2\beta(1-\beta)(\bar{Y}_a+\bar{Y}_b-\bar{Y}_{ab})^2}{[r_i+\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2}} \right\}$$

where $r = \frac{n_A}{n_B}$. Generally, only a few iterations are required to obtain r starting from a reasonable "guess" for r_1. The estimator and its variance are not sensitive to deviations from r_0 (optimum) of 10 percent or less. An estimator of the optimum p (i.e., p_0) from the sample data is:

$$\hat{p} = \frac{\frac{N_A n_a}{2}\bar{y}_a + \frac{N_B n_b}{2}(\bar{y}_{ab}-\bar{y}_b)}{\frac{n_A}{n_B}}$$
$$\frac{\frac{N_A n_a}{2}}{n_A} + \frac{\frac{N_B n_b}{2}}{n_B})\bar{y}_{ab}$$

But $\hat{p}$ is now a function of several sample statistics which disturbs the unbiasedness of the estimator. However, the degree of bias is considered to be negligible. An alternative estimator of p is available, but requires the parameter σ_a^2, σ_{ab}^2 and σ_b^2. This is the bi-quadratic solution given by Hartley.

5.3.5 Assumption of Equality of Means for "Overlap" Domains

In practice, we face the problem of pooling of independent estimates of the parameter $\hat{Y}_{ab}$ or $\bar{Y}_{ab}$ from different frames. Each estimate is given with its sample size and estimated standard error. Can the estimates be considered as homogeneous? That is, are they estimating the same quantity? Let $n = n_1+\ldots+n_K$ equal the samples corresponding to each frame and denote by π_i the ratio n_i/n. The asymptotic distribution of $\sqrt{n_i}\,(T_i-\Theta_i)$ is $N[0,S_i^2(\Theta_i)]$.

Consider, $$H = \sum^{K} \frac{n_i (T_i - \hat{\Theta})^2}{S_i^2(T_i)} = n \sum^{K} \frac{\pi_i (T_i - \hat{\Theta})^2}{S_i^2(T_i)}$$

where T_i is the estimate of the parameter Θ from the i^{th} frame, and $\hat{\Theta}$ is given by

$$\hat{\Theta} = \frac{\sum^{K} \pi_i T_i}{S_i^2(T_i)} \div \sum^{K} \frac{\pi_i}{S_i^2(T_i)}$$

H is distributed as χ^2 with (K-1) degrees of freedom as $n \to \infty$. If $H > \chi^2_{\alpha, K-1}$, the equality of the means is rejected and the estimates should not be pooled.

5.3.6 The Special Case of Frame A With 100 Percent Coverage

If Frame A is complete (covers all the units in the population) then $N_A = N$, $N_{ab} = N_B$, $N_a = N_A - N_B$, $N_b = 0$, so we are in case 2. Since $N_a = N_A - N_B > 0$, Frame B must have fewer units than Frame A.

5.3.7 Different Units in Frames With Overlapping Characteristics

In this case, the elementary units which make up the frame are different. Consider a survey in a city to estimate the total cost expended on the laundering of clothes; both private households and commercial laundries will launder items which we refer to as "clothes." A portion of "clothes" belonging to a household may be sent to a laundry and the rest washed in the home. A commercial laundry handles clothes from households and from some "commercial institutions" which send all their laundry out. That is, the characteristic pertaining to the elementary unit is partitioned rather than assigning the unit to either domain (a), (ab), or (b). The three domains are: (1) household clothes laundered in the home, (2) household clothes laundered in commercial laundries, and (3) commercial institution clothes laundered in commercial laundries. The characteristic of interest might be dollars spent or pounds of clothes, or both.

For each frame the characteristic of interest is defined as follows:

Frame A $_{j}y_{i}$ = $\begin{cases} {}_{j}y_{i} & \text{if the clothes in the } i^{th} \text{ home are laundered} \\ & \text{in the home } (j^{th} \text{ domain} = a) \\ {}_{qv}y_{i} & \text{if the clothes in the } i^{th} \text{ home are laundered in a} \\ & \text{commercial laundry (i.e., } j^{th} \text{ domain} = ab) \end{cases}$

Frame B $_{j}y_{k}$ = $\begin{cases} {}_{r}y_{k} & \text{if clothes in the } k^{th} \text{ commercial laundry are from} \\ & \text{commercial institutions } (j^{th} \text{ domain} = b) \\ {}_{qv}y_{k} & \text{if clothes in the } k^{th} \text{ commercial laundry are from} \\ & \text{a home } (j^{th} \text{ domain} = ab) \end{cases}$

The unbiased estimate of the population total is given by

$$\hat{Y} = \frac{N_a}{n_A} \left[\sum_{a} {}_{j}y_{i} + p \sum_{ab} {}_{j}y_{i} \right] + \frac{N_B}{n_B} \left[\sum_{b} {}_{j}y_{K} + q \sum_{ab} {}_{qv}y_{K} \right]$$

$$V(\hat{Y}) = \frac{N_A^2}{n_A} \left(1 - \frac{n_A}{N_A}\right) S^2_{{}_{j}y_{i}} + \frac{N_B^2}{n_B} \left(1 - \frac{n_B}{N_B}\right) S^2_{{}_{j}y_{K}}$$

and the sample estimator of the variance is a copy of $V(\hat{Y})$.

5.3.8 Exercises 11, 12, and 13

These three exercises are for a two-frame survey for a single stratum in a State in which the area frame (Frame A) provides access to 100 percent of farm operations and the list frame (Frame B) provides access to an unknown fraction of the farm operations. The names have been coded by replacing the surname of the person, or principal name of a joint operation, or business firm by the permutation of three letters of the alphabet. Each different ordering of the three letters represents a different surname or principal business name; that is, ABC is different than BAC. The city name in the address has likewise been coded by numbers from 01 to 99. If the city was omitted by error, the code is given as 00. The agricultural land operations are associated with the area frame units only

through the land inside the segment. Economic units such as "farms" are associated with an operator who must reside in one and only one segment during the survey period. Agricultural land and economic units are associated with name and address units (or combinations) for the list frame.

Exercise 11:

Each student is to determine units overlapped between frames. From the listings of respondents for the area frame (Frame A listing) and list frame (Frame B listing), determine the operations which can be accessed by both frames (i.e., overlap domain). After the last page of the listings, the operations which were identified as in the overlap domain by the survey staff are given.

Exercise 12:

Each student is to determine duplication within each frame. Using the listings given for exercise 11, determine which (if any) respondent operations are duplicated within the area frame and also any duplications within the list frame. The duplications determined by the survey staff during the survey are given at the end of the listings in exercise 11.

Exercise 13:

Each student is to calculate the survey estimates (i.e., totals). Table 4 shows the survey means, sample sizes, and expansion factors by frames which are used to obtain estimates by domains and the total population.

(1) Compute the two estimates for the overlap domain. Should these two estimates be pooled and used in either the screening or multiple frame estimator? (See 5.3.5.)

(2) Compute and compare the estimates obtained by using the following three estimators of the total population:

a. Area frame estimator ($p = 1$, $q = 0$).

b. Screening estimator ($p = 0$, $q = 1$).

c. Multiple frame estimator (use weights given in (3), page 102, based on the number of frame units selected for the sample where $p, q \geq 0$; $p + q = 1$).