Bureau of the Census

Draft

Survey methods and sampling

1999

CHAPTER 1

GENERAL NATURE OF SAMPLE SURVEYS

1.1 ROLE OF SAMPLING IN STATISTICAL THEORY AND METHODS

In a broad sense, sampling theory can be considered as coextensive with modern statistical methods. Almost all of the modern developments in statistics relate to the inferences that can be made about a population when information is available from only a sample of the elements of the population. Some of the ways in which this is reflected in statistical programs are mentioned below.

1.1.1 Survey Work

In most survey work, the population consists of all persons (or housing units, households, industrial establishments, farms, etc.) in a city or other area. Information is obtained or desired from a sample of the population, but inferences are required on characteristics of the whole population.

1.1.2 Design and Analysis of Experiments

In the design and analysis of experiments, the population represents all possible applications of several alternative techniques which can be used. For example, the experiment may be agricultural, in which a number of fertilizers are being tested. The population is infinite because it represents the use of the fertilizers in all possible farms over all time. The problem is to design experiments so that the maximum amount of information can be made available for inferences about the full population, estimated from a sample of limited size.

1.1.3 Quality Control

In the application of quality control methods in an industrial establishment, for example, the population is all of the products coming out of a machine. Inferences are needed on how well the products conform to specifications. The term "quality control" is also applied to a sample check on the quality of field work done in a sample survey; the sample check is carried out after the actual survey is completed. Office operations such as editing and coding are also subject to quality control; a sample of the work is checked to determine if it meets acceptable standards.

1.2 CONTENTS OF CHAPTERS

These chapters will be limited to one aspect of sampling; that is, sampling application in survey work. They will deal mainly with principles of sampling from the common sense rather than the mathematical viewpoint, though mathematics cannot be entirely avoided. The emphasis will be on the methods of sampling that can be used under different conditions. The formulas will be presented, some without mathematical proof, but with information on how they should be used. Two types of examples will be used to illustrate the formulas and methods: (a) simple examples to make the techniques clear, and (b) examples taken from actual surveys to show the realistic applications of the methods discussed.

First there will be a general discussion of the subject as a whole, including the nature of probability sampling, and choices of sampling units and sampling frames. Then we shall describe the types of common sample designs--simple random sampling, stratified sampling, and cluster sampling. The features of these designs and the methods of sample selection will be discussed. The different methods of estimating the characteristics of the population from the sample results will also be treated, as well as how to determine the size of sample required for a particular degree of reliability and how to calculate sampling errors.

We shall also discuss the problem of estimating, from a sample, the results that would have been obtained from a full census using the same questionnaire, enumeration or interview procedures, supervision, etc. These are aspects of the problem of sampling error. There are, of course, nonsampling errors that arise from wrong responses to questions, or from poorly worded questions. These are present in complete censuses as well as in sample surveys. Although the lectures are not primarily concerned with such nonsampling errors, they may be very important. In fact, nonsampling errors often represent more serious limitations on the use of statistics than sampling errors.

1.3 REASONS FOR THE USE OF SAMPLES

There are six basic reasons for the use of samples:

(1) A sample may save money (as compared with the cost of a complete census) when absolute precision is not necessary.

(2) A sample saves time, when data are desired more quickly than would be possible with a complete census.

(3) A sample may make it possible to concentrate attention on individual cases.

(4) In industrial uses, some tests are destructive (for example, testing the length of time an electric bulb will last) and can only be performed on a sample of items.

(5) Some populations can be considered as infinite, and can, therefore, only be sampled. A simple example is an agricultural experiment for testing fertilizers. In one sense, a census can be considered as a sample at one instant of time of an underlying causal system which has random features in it.

(6) Where nonsampling errors are necessarily large, a sample may give better results than a

complete census because nonsampling errors are easier to control in smaller-scale operations.

1.4 ILLUSTRATIONS OF SAMPLING

The following illustrate the use of sampling in various situations.

1.4.1 Limited Funds

The use of a sample survey when limited funds are available for collecting information is well known. Sampling may also be used to save money in tabulation. For example, in the 1950 Census in the United States most of the data were collected on a 100-percent basis. However, many tabulations were made on a sample basis (20% or 3-1/3%) for special detailed classifications to save the cost of tabulating 150,000,000 individual records. The 1960 Census utilized sampling procedures to an even greater extent in both the collection and the tabulation of data.

1.4.2 Time Saving

Other examples from the 1950 census in the United States illustrate how samples can be used to save time. The enumeration of the census was taken in April 1950. The time required for processing the results was such that publication of the results was expected to start in 1951 and continue through 1952. A sample of the census results was selected for quick processing and tabulation, and preliminary results were published on the basis of this sample. These results were issued 1 to 2 years earlier than the complete census results.

1.4.3 Concentration on Particular Cases

Some surveys require such intensive and time-consuming interviews that it is impossible to consider them on any basis except a sample basis. Moreover, the use of sampling permits particular attention to be given to a limited number of cases. Examples are family budget studies and comprehensive studies of health conditions.

1.4.4 Sampling for Time Series

Information may be required for a time series when data are available only for particular periods of time and results are needed promptly. The series may be one of economic activity in the country, with figures available only on a yearly or monthly basis, or it may be one of producing a learning curve for which only occasional tests are possible.

1.4.5 Controlling Nonsampling Errors

An interesting example arose in the 1950 United States Census of a case where the relationship between nonsampling and sampling errors made sample results preferable to complete census results. The United States has conducted a monthly sample survey of the labor force since 1940. In 1950, it was based on a sample of 20,000 households. The information obtained in the 1950 complete census also included labor force status. When the results of the census became available, it was clear that the figures for both unemployed and employed persons were quite

different from those estimated from the labor force sample survey; the differences were far beyond what could be expected on the basis of the sampling errors. The problem of reporting in the census introduced much greater error than the sampling error of the monthly survey (this greater error was caused by the use of enumerators who, for the most part, were inexperienced in interviewing). Users of census data were advised, therefore, to use the sample results as the more reliable national statistics on the labor force.

1.5 LIMITATIONS OF SAMPLING

Under certain conditions, the usefulness of sampling becomes questionable. Three principal conditions can be mentioned.

(1) If data are needed for very small areas, disproportionately large samples are required, since precision of a sample depends largely on the sample size and not on the sampling rate. In this case, sampling may be almost as expensive as a complete census.

(2) If data are needed at regular intervals of time, and it is important to measure very small changes from one period to the next, very large samples may be necessary.

(3) If there are unusually high overhead costs connected with a sample survey, caused by work involved in sample selection, control, etc., sampling may be impractical. For example, in a country with many small villages it may be more economical to enumerate all the households in the sample villages than to enumerate a sample of households within the sample villages. For office processing, however, a sample of the enumerated households may be used to reduce the work and costs of producing tabulations.

CHAPTER 2

CRITERIA AND DEFINITIONS

2.1 CRITERIA FOR THE ACCEPTABILITY OF A SAMPLING METHOD

It has been demonstrated repeatedly in practical applications that modern sampling methods can provide data of known reliability on an efficient and economical basis. However, although a sample includes only part of a population, it would be misleading to call a collection of numbers a "sample" merely because it includes part of a population.

To be acceptable for statistical analysis, a sample must represent the population and must have measurable reliability. In addition, the sampling plan should be practical and efficient.

2.1.1 Chance of Selection for Each Unit

The sample must be selected so that it properly represents the population that is to be covered. This means that each unit (farm, household, person, or whatever unit is being sampled) must have a **nonzero probability** (chance) of being selected.

2.1.2 Measurable Reliability

It should be possible to measure the reliability of the estimates made from the sample. That is, in addition to the desired estimates of characteristics of the population (totals, averages, percentages, etc.) the sample should give measures of the precision of these estimates. As we shall see later, these measures of precision can be used to indicate the maximum error that may reasonably be expected in the estimates, if the procedures are carried out as specified, and if the sample is moderately large. The estimation of precision is not possible unless the selection is carried out so that the chance of selection of each unit is known in advance and random sampling is used.

2.1.3 Feasibility

A third characteristic is that the sampling plan must be practical. It must be sufficiently simple and straightforward so that it can be carried out substantially as planned; that is, the sampling theory and practice will be the same. A plan for selecting a sample, no matter how attractive it may appear on paper, is useful only to the extent that it can be carried out in practice. When the methods actually followed are the same (or substantially the same) as specified in the sampling plan, then known sampling theory provides the necessary measures of reliability. In addition, the measures of reliability computed from the survey results will serve as powerful guides for future improvement in important aspects of the sample design.

2.1.4 Economy and Efficiency

Finally, the design should be efficient. Among the various sampling methods that meet the three criteria stated above, we would naturally choose the method which, to the best of our knowledge, produces the most information at the smallest cost. Although this is not an essential feature of an acceptable sampling plan, it is clearly a highly desirable one. It implies that the most effective possible use will be made of all available facilities and resources, such as maps, other statistical data, personal knowledge, sampling theory, etc.

We shall consider only sampling methods that conform to the above criteria. We shall present basic theory for various alternative designs which are possible, and methods of measuring their precision. We shall also stress practical methods of application and considerations of efficiency.

2.2 DEFINITIONS OF TERMS

2.2.1 Statistical Survey

The statistical survey is an investigation involving the collection of data. Observations or measurements are taken on a sample of elements for making statistical inferences (see Glossary in Annex A) about a defined group of elements. Surveys are conducted in many ways.

2.2.2 Unit of Analysis

The **unit of analysis** is the unit for which we wish to obtain statistical data. The most common units of analysis are persons, households, farms, and business firms. They may also be products coming out of some machine process. The unit of analysis is frequently called an element of the population. There may be more than one unit of analysis in the same survey; for example, households and persons; or number of farms and hectares (or acres) harvested.

2.2.3 Characteristic

A **characteristic** is a general term for any variable or attribute having different possible values for different individual units of sampling or analysis. In a sample survey, we observe or measure the values of one or more characteristics for the units in the sample. For example, we observe (or ask about) the area of land for rice crop, the number of cattle on a farm, the age and sex of a person, the number of children per family, etc. So, we observe a unit, but we measure several characteristics of that unit.

2.2.4 Population or Universe

The **population or universe** is the entire group of all the units of analysis whose characteristics are to be estimated. The chapters in this sampling manual will deal primarily with a finite population, having N units.

2.2.5 Probability Sample

A probability sample is a sample obtained by application of the theory of probability. In probability sampling, every element in a defined population has a known, nonzero, probability of being selected. It should be possible to consider any element of the population and state its probability of selection.

2.2.6 Sampling with Replacement and Without Replacement

A simple way of obtaining a probability sample is to draw the units one by one with a known probability of selection assigned to each unit of the population at the first and each subsequent draw. The successive draws may be made with or without replacing the units selected in the preceding draws. The former is called the procedure of sampling with replacement, and the latter, sampling without replacement.

2.2.7 Simple Random Sampling

Simple random sampling is a special case of probability sampling, sometimes called unrestricted random sampling. It is a process for selecting n sampling units one at a time, from a population of N sampling units so that each sampling unit has an equal chance of being in the sample. Every possible combination of n sampling units has the same chance of being chosen. Selection of one sampling unit at a time with equal probability may be accomplished by either sampling with replacement or without replacement. Almost, if not all, samples are selected without replacement. Using a table of random numbers to select the units satisfies this definition of simple random sampling.

2.2.8 Sampling Frame

The totality of the sampling units from which the sample is to be selected is called the **sampling frame**. The frame may be a list of persons or of housing units; it may be a subdivided map, or it may be a directory of names and addresses stored in some kind of electronic medium, such as a file in a hard disk or a data base.

2.2.9 Parameter

A **parameter** is a quantity computed from all values in a population set. That is, a parameter is a descriptive measure of a population. For example, consider a population consisting of N elements. Then the population total, the population average or any other quantity computed from measurements including all elements of the population is a parameter. ***The objective of sampling is to estimate the parameters of a population***.

2.2.10 Statistic

A **statistic** is a quantity computed from sample observations of a characteristic, usually for the purpose of making an inference about the characteristic in the population. The characteristic may be any variable which is associated with a member of the population, such as age, income, employment status, etc.; the quantity may be a total, an average, a median, or other quantiles. It may also be a rate of change, a percentage, a standard deviation, or it may be any other quantity whose value we wish to estimate for the population.

Note that the term **statistic** refers to a sample estimate and the term **parameter** refers to a population value.

Note on Quantiles: What is a quantile? If a set of data is arranged in order of magnitude, the middle value (or the arithmetic mean of the two middle values) which divides the set into two equal parts is the MEDIAN. By extending this idea we can think of those values which divide the set into four equal parts. These values, denoted by Q_1, Q_2 and Q_3 are called the first, second and third *quartiles* respectively, the value of Q_2 being equal to the median. Similarly the values which divide the data into ten equal parts are called *deciles* and are denoted by D_1, D_2, ... D_9, while the values dividing the data into one hundred equal parts are called *percentiles* and are denoted by P_1, P_2, ... P_{99}. The 5th decile and the 50th percentile correspond to the median. The 25th and 75th percentiles correspond to the first and third quartiles, respectively. Collectively, quartiles, deciles, percentiles and other values obtained by equal subdivisions of the data are called *quantiles*.

2.2.11 Independent Information

Independent information consists of data that are known in advance of or simultaneously with the survey which are not based on the survey but are used to improve the survey design. Such data may be used for purposes of stratification, for determining the probabilities of selection, or in estimating the final results from the sample data. The data must be of good, known quality.

2.2.12 Estimate and Estimator

An **estimate** is a numerical quantity computed from sample observations of a characteristic and intended to provide information about an unknown population value.

An **estimator** is a mathematical formula or rule which uses sample results to produce an estimate for the entire population. For example, the sample average,

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

is an estimator. It provides an estimate of the parameter, the population average,

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

that is, the sample average is an estimate of the population average.

Therefore, the estimator refers to a mathematical formula. When numbers are plugged into the formula, an estimate is produced. However, in common statistical language, the words estimate and estimator are used interchangeably.

2.2.13 Probability of Selection

The **probability of selection** is the chance that each unit in the population has of being included in the sample. Probability values range from 0 to 1, inclusive.

2.2.14 Random Variables

A **random variable** is a variable which, by chance, can be equal to any value in a specified set. The probability that it equals any given value (or falls between two limits) is either known, can be determined, or can be approximated or estimated. A chance mechanism determines the value which a random variable takes. For example, in flipping a coin, we can define the random variable X which can take the value 1 is the coin lands ‘heads’ and the value 0 if the coin lands ‘tails’. Therefore, the variable X, as was just defined, can take either one of two values after the coin is flipped.

2.2.15 Probability Distribution

The **probability distribution** gives the probabilities associated with the values which a random variable can equal. If there are N values that a random variable X can take, say X_1, X_2, ... ,X_N, then there are N probabilities associated with the $X_{i's}$ values, namely P_1, P_2, ... ,P_N. The probabilities and the values the random variable takes constitute the probability distribution of X.

2.2.16 Illustration

The 1980 U.S Census of Population and Housing found that 217,482,000 persons lived in 79,108,000 households and that 8,958,000 persons lived in institutions and other group quarters[1]. Table 2.1 below shows the distribution of households of different sizes.

These data show that 22.5% of all U.S. households contain just one person; 31.3% are two-person households, and so on. Now if we were to pick a household at random, what is the probability that we would pick a one-person household? If each household, large or small, is equally likely to be picked, then there is a .225 probability of picking a one-person household.

[1] *The Census Bureau defines a household as persons who occupy a house, apartment, or other separate living quarters. One of the tests in determining a household is that there are complete kitchen facilities for the exclusive use of the occupants. People who are not in households live in group quarters including rest homes, rooming houses, military barracks, jails, and college dormitories.*

Table 2.1[2].

SIZE OF U.S. HOUSEHOLDS, 1980

SIZE	NUMBER OF HOUSEHOLDS	FRACTION OF TOTAL HOUSEHOLDS
1 Person	17,816,000	.225
2 Persons	24,734,000	.313
3 Persons	13,845,000	.175
4 Persons	12,470,000	.158
5 Persons	5,996,000	.076
6 Persons	2,499,000	.032
7 or more	1,748,000	.022
TOTAL	79,108,000	1.00

[2] *Source: U.S. Bureau of the Census, Current Population Survey in Statistical Abstract of the United States. (Washington, D.C. : U.S. Government Printing Office, 1981)*

Study Assignment

Exercise 1. *In order to select a sample of the total population of a city, a sample is selected from the telephone directory for that city and the families of the persons selected are interviewed. Does this satisfy the criteria for acceptability? Explain.*

Exercise 2. *In order to determine the population of a city where all children of school age attend school, a sample of school children is drawn and their families are interviewed. Give two reasons why this does not meet the criteria for acceptability. (Think of families who have more than one child in school and families that don't have any children.)*

Exercise 3. *Suppose that you were using sampling to estimate the total number of words in a book that contains illustrations.*

(a) Is there any problem of definition of the population?

(b) What are the pros and cons of (1) using the page, (2) the line as a sampling unit?

Exercise 4. *Suppose that you work for a major public opinion pollster and you wish to estimate the proportion (See Glossary in Annex A) of adult citizens who think the President is doing a good job in heading the nation's economy. Clearly define the population you wish to sample.*

Exercise 5. *The problem of finding a frame that is complete and from where a sample can be drawn is often an obstacle. What kinds of frames might be tried for the following surveys? Do the frames have any serious weakness?*

(a) A survey of stores that sell luggage in a large city.

(b) A survey of the kinds of articles left behind in subways or on buses.

(c) A survey of persons bitten by snakes during the last year.

(d) A survey to estimate the number of hours per week spent by family members watching television.

CHAPTER 3

SIMPLE RANDOM SAMPLING
SAMPLING DISTRIBUTION

3.1 INTRODUCTION

In this chapter, we shall introduce the concept of the sampling distribution of a statistic, probably the most basic concept of statistical inference. We shall concentrate only on the sample mean and its sampling distribution. We shall first introduce certain definitions and relationships of terms needed for the sampling distribution.

3.2 EXPECTED VALUE

The expected value is the average value for a single characteristic over all possible samples. Mathematically, we define the expected value (or mean) of a random variable Y as follows:

$$E(y) = \sum_{all\ y} y.p(y)$$

where $\sum_{all\ y} p(y) = 1$ and the Σ Greek letter is used to indicate the sum of the products of possible values of y and their associated probabilities p(y). The small y denotes a particular value of Y.

The expected value is a weighted average of the possible outcomes, with the probability weights reflecting the likelihood of occurrence of each outcome. Thus, the expected value should be interpreted as the long-run average value of Y, if the frequency with which each outcome occurs is in accordance with its probability.

For example, consider Table 2.1 in which the random variable Y is used to represent the size of a U.S. household selected at random. We write the expected value of Y as:

$$E(y) = (1)\ (.225) + (2)\ (.313) + (3)\ (.175) + (4)\ (.158) + (5)\ (.076) + (6)\ (.032) + (7.7)\ (.022) = 2.75$$

The expected value of Y is not the most likely or the most typical value of Y. It is the long-run average value of Y, if we repeatedly select households at random. Some households have fewer than 2.75 people; some have more. The average of these different household sizes is 2.75.

Note that the category "7 or more" aggregates data for households where Y = 7, 8, 9,...; so, it would be misleading to use Y = 7. Instead, we have put this .022 probability at Y = 7.7, which is the average size of households with seven or more persons.

3.2.1 Unbiased Estimate

A type of estimate having the property that the average of such estimates obtained from all possible samples of a given size is equal to the true value. Mathematically, an estimate is unbiased if the expected value of the estimate is equal to the parameter being estimated.

For example, if $\hat{\theta}$ is an estimate of the parameter θ and if $E(\hat{\theta}) = \theta$ $\hat{\theta}$ then is an unbiased estimate of θ. Otherwise, $E(\hat{\theta}) - \theta = \mathit{Bias}.$ That is, the bias is the difference between the expected value of an estimate and the true population value (parameter) being estimated.

3.2.2 Consistent Estimate

An estimate is consistent if its values tend to concentrate increasingly around the true value as the sample size increases. In other words, the estimate assumes the population value with probability approaching unity as the sample size tends to infinity. This definition of consistency strictly applies to estimates based on samples drawn from an infinite population. We use the following definition in the case of a finite population. An estimate $\hat{Y}$ is said to be a consistent estimate of the parameter Y if it takes the population value when n=N.

In the next section we will see that for simple random sampling the sample mean is an unbiased and consistent estimate of the population mean as the sample size increases.

3.3 SAMPLING DISTRIBUTION

A sampling distribution is the probability distribution of all possible values that an estimate might take under a specified sampling plan.

In this section we will show by examples that the sample average (mean) is both an unbiased and a consistent estimate of the true population average.

Let us first present the idea of a sampling distribution of the mean by actually listing all possible random samples of size n=2 which can be drawn from a hypothetical population of N=5 housing units (HUs) shown in Table 3.1. We wish to estimate the average household (HH) size of these HUs from a sample.

Table 3.1 HOUSEHOLD SIZE PER HOUSEHOLD

HU (U_i)	HH SIZE (Y_i)
U_1	3
U_2	5
U_3	7
U_4	9
U_5	11

The total number of persons in the population is:

$$Y = \sum_{i=1}^{N} Y_i = 35$$

The average number of persons per household (or average household size) is:

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{35}{5} = 7$$

If we take a sample of size 2 from this population, there are $\binom{5}{2}$ =10 possibilities, and they are:

3 and 5, 3 and 7, 3 and 9, 3 and 11

5 and 7, 5 and 9, 5 and 11

7 and 9, 7 and 11

9 and 11

The means of these samples are 4, 5, 6, 7, 6, 7, 8, 8, 9, and 10, respectively, and if sampling is random so that each sample has the probability 1/10, we obtain all the possible samples of size two HUs from a population of 5 HUs, as shown in Table 3.2. Table 3.3 presents the sampling distribution of the mean.

TABLE 3.2

SAMPLES OF TWO HUs FROM A POPULATION OF 5 HUs.

SAMPLES OF SIZE n = 2	VALUE OF $\bar{y}$	PROBABILITY p(y)
3,5	4	1/10
3,7	5	1/10
3,9	6	1/10
3,11	7	1/10
5,7	6	1/10
5,9	7	1/10
5,11	8	1/10
7,9	8	1/10
7,11	9	1/10
9,11	10	1/10

TABLE 3.3

SAMPLING DISTRIBUTION OF THE MEAN.

MEAN $\bar{y}$	PROBABILITY p(y)
4	1/10
5	1/10
6	2/10
7	2/10
8	2/10
9	1/10
10	1/10

An examination of this sampling distribution reveals some pertinent information relative to the problem of estimating the mean of the given population using a random sample of size 2. For instance, we see that corresponding to $\bar{y}$ = 6, 7, or 8, the probability is 6/10 that a sample mean will not differ from the population mean (which is 7) by more than 1, and that corresponding to $\bar{y}$ = 5, 6, 7, 8, or 9, the probability is 8/10 that a sample mean will not differ from the population mean by more than 2.

Further useful information about this sampling distribution of the mean can be obtained by calculating its expected value as follows:

$$E(\bar{y}) = (4)\frac{1}{10} + (5)\frac{1}{10} + (6)\frac{2}{10} + (7)\frac{2}{10} + (8)\frac{2}{10} + (9)\frac{1}{10} + (10)\frac{1}{10} = 7$$

We may also use Table 3.2 to compute the expected value of $\bar{y}$:

$$E(\bar{y}) = \frac{1}{10}\sum_{1}^{10} \bar{y} = \frac{70}{10} = 7 = \bar{Y}$$

Note that the same results would be obtained for samples of any size. Recall the definition of the expected value, which is the average of a single characteristic over all possible samples.

With simple random sampling the sample mean is an unbiased estimate of the true mean.

We will now compare the distribution of the sample estimates to show that:

(1) As the sample size increases, the means of the samples tend to concentrate more and more around the true average value. In other words, the estimates tend to become more and more reliable as the sample size increases.

(2) The percentage distributions of the sample estimates can be used to predict the chance of obtaining a sample estimate within specified ranges of the true value. To see the above statements, consider a hypothetical population of 12 individuals. We wish to make different estimates from a sample of 1,2,3,4,5,6 and 7 individuals. The full population is shown in the Table 3.4 below.

Table 3.4

INCOMES OF HYPOTHETICAL POPULATION OF 12 PERSONS

INDIVIDUAL	INCOME
A	$1,300
B	6,300
C	3,100
D	2,000
E	3,600
F	2,200
G	1,800
H	2,700
I	1,500
J	900
K	4,800
L	1,900
TOTAL INCOME	$32,100
AVERAGE INCOME	$2,675

A frequency distribution of the sample means is illustrated in Table 3.5 for samples of 1,2,3,4,5,6 and 7 individuals. For each sample size, the percentage of the sample estimates falling within a specified range of the true value and the average of the means are also shown in the table.

For example, the proportion of the sample results falling between $2,000 and $3,400 is 47% for samples of 2; 58% for samples of 3; 69% for samples of 4; and 78%, 87% , and 94% for samples of 5,6, and 7 respectively. This tells us that by taking samples large enough, the proportion of the sample estimates falling within a designated interval about the expected value can be made as close to 100% as desired. That is, we can predict the precision of a sample if we have the distribution of all sample estimates of a given size for the population. The increasing concentration of sample estimates around the true value illustrates consistency, a quality possessed by important types of sample estimates.

TABLE 3.5

ALL POSSIBLE ESTIMATES OF AVERAGE INCOME
FROM SAMPLES DRAWN WITHOUT REPLACEMENT
FROM THE POPULATION OF 12 PERSONS

Average Income Estimated from Sample	Number of samples having indicated estimate of average income with sample of size n						
	n =1	n = 2	n = 3	n = 4	n = 5	n = 6	n= 7
$ 800 to $1,199	1	1	-	-	-	-	-
$1,200 to $1,399	1	2	3	1	-	-	-
$1,400 to $1,599	1	5	10	11	7	1	-
$1,600 to $1,799	-	6	15	25	25	16	6
$1,800 to $1,999	2	5	20	42	55	50	27
$2,000 to $2,199	1	6	22	50	78	84	61
$2,200 to $2,399	1	6	22	52	90	109	98
$2,400 to $2,599	-	6	19	52	101	139	136
$2,600 to $2,799	1	3	17	49	108	151	150
$2,800 to $2,999	-	4	16	57	101	133	130
$3,000 to $3,199	1	3	16	46	81	107	108
$3,200 to $3,399	-	3	16	38	61	79	62
$3,400 to $3,599	-	2	13	26	46	43	14
$3,600 to $3,799	1	2	10	21	27	12	-
$3,800 to $3,999	-	3	7	11	10	-	-
$4,000 to $4,199	-	3	4	10	2	-	-
$4,200 to $4,399	-	2	6	3	-	-	-
$4,400 to $4,599	-	1	1	1	-	-	-
$4,600 to $4,799	-	1	2	-	-	-	-
$4,800 to $6,399	2	2	1	-	-	-	-
Number of Samples	12	66	220	495	792	924	792
Average of all possible samples·	$2,675	$2,675	$2,675	$2,675	$2,675	$2,675	$2,675

· Expected Value

This means that if the sample is sufficiently large, one takes very little risk in using sample estimates. (From the above illustration, it might appear that the increase in concentration arises from the fact that, as the size of the sample increases, the percentage of the population in the sample becomes higher. Actually, similar results would be observed when the size of sample increases even though only a small proportion of the universe is included.)

3.4 PREDICTING RELIABILITY OF SAMPLE ESTIMATES (CONFIDENCE INTERVAL)

We have seen that the precision of a sample can be predicted if we have the distribution of all sample estimates of a given size for the population. In a real situation, we can not select all possible samples and examine the estimates derived from them. We must depend upon a single sample. Therefore, it is necessary to find some measure of the extent to which the estimates made from various samples differ from the true value; this measure, if it is to be useful, must be one that can be estimated from the sample itself. Before showing how and why we can do this, we shall introduce certain definitions and relationships which are derived from the theory of sampling.

3.4.1. Standard Deviation

We shall show that there is a measure of the variability in the original population which can be estimated from the observations in a single sample, and from which it is possible to estimate the expected error in the sample mean.

The measure of variability in the population is called the **standard deviation**; its square is called the **population variance** and is designated by the symbol σ^2 or VAR. The variance of the population is defined as the average of the squares of the deviations of all the individual observations from their mean value. Thus, it would be computed by the following process, if all the values in the universe could be observed:

$$\sigma^2 = \frac{(Y_1-\bar{Y})^2+(Y_2-\bar{Y})^2+...+(Y_N-\bar{Y})^2}{N} = \frac{1}{N}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$$

where the Y's with subscripts are individual observations and $\bar{Y}$ is the mean of the N observations for the N elements in the universe. Note that it has become fairly general practice to denote the population variance by σ^2 when dividing by N, and by S^2 when dividing by N-1; symbolically,

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$$

Its sample equivalent is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

where n is the sample size, y_i is the sample measurement of a characteristic and $\bar{y}$ is the sample mean.

We will use S^2 throughout the text because s^2 is an unbiased estimate of S^2. Note that all results are equivalent in either notation. Also,

$$\sigma^2 = \frac{N-1}{N}S^2\ , \quad S^2 = \frac{N}{N-1}\sigma^2$$

3.4.2 Standard Error of Sample Means

The variance of the sample means is the average of the squares of the deviations of the means of all possible samples of size n from the true mean. The variance of $\bar{y}$ is denoted $S^2(\bar{y})$ and we write:

$$S^2(\bar{y}) = \frac{S^2}{n}(\frac{N-n}{N})$$

The square root of the variance of $\bar{y}$ is called the **standard error** for means of samples of size n.

The standard error of $\bar{y}$ is:

$$S(\bar{y}) = \frac{S}{\sqrt{n}}\sqrt{(\frac{N-n}{N})}$$

It is important to note that the standard error varies with the size of the sample, as we would expect. If we compute the standard error for all possible samples of sizes shown in Table 3.5, we see that as the sample size increases, the standard error becomes smaller and smaller. This is shown in the following illustration (see Table 3.6). The factor $(\frac{N-n}{N})$ in the formula for the variance of $\bar{y}$ is called the **finite population correction** (fpc). As a rule of thumb, if $n \leq 0.05N$ we can ignore $(\frac{N-n}{N})$, since its value will be close to 1. Otherwise we should include it in the formula in order not to severely overestimate the variance of $\bar{y}$.

3.4.3 Illustration

Consider again the population of 12 individuals and Table 3.4. In this case, the true average is $\bar{Y}$ = \$2,675 with N=12. We compute S^2 as follows:

$$S^2 = \frac{(1{,}300-2{,}675)^2+(6{,}300-2{,}675)^2+......+(1{,}900-2{,}675)^2}{11} = \$2{,}469{,}318.18$$

and S = \$1,571.41

Using S, we can compute the standard error of the sample mean $(\bar{y})$ for different sample sizes n: For example, if the sample size n=1 then,

$$S(\bar{y}) = \frac{1571.41}{\sqrt{1}} \sqrt{\left(\frac{12-1}{12}\right)} = \$1,505$$

for n = 2,

$$S(\bar{y}) = \frac{1571.41}{\sqrt{2}} \sqrt{\left(\frac{12-2}{12}\right)} = \$1,015$$

The standard errors for all possible sample sizes are given in the following table.

Table 3.6

STANDARD ERROR OF ESTIMATES OF AVERAGE INCOME
FOR VARIOUS SAMPLE SIZES

SIZE OF SAMPLE	STANDARD ERROR OF ESTIMATED MEASURE (Y)
1	\$1,505
2	1,015
3	786
4	642
5	537
6	454
7	383

3.4.4 Interval Estimate (Confidence Interval)

We know that the probability of an estimate being equal to the true value (parameter) is zero for continuous variables. Thus, it will be more useful if we can state how probable it is that an interval based on our estimate will contain the parameter to be estimated.

Interval estimator - An interval estimator is a formula that tells us how to use the sample observations to calculate two numbers that define an interval which will enclose the estimated parameter with a certain (usually high) probability. **The resulting interval is called a confidence interval and the probability that it contains the true parameter is called its confidence coefficient. If a confidence interval has a confidence coefficient equal to .95, we call it a 95% confidence interval.**

In general, the confidence interval for a parameter θ is given by $[\hat{\theta} \pm tS(\hat{\theta})]$.

The symbol t is the value of the normal deviate corresponding to the desired confidence probability. In practice, S^2 is not known. Usually, s^2, the sample variance (see Chapter 4) is calculated from the sample data and used as an estimate of S^2. If n is large, s provides a fairly good estimate of S; however, for small samples this may not be the case. Using s, the confidence interval is $[\hat{\theta} \pm ts(\hat{\theta})]$.

For the parameter $\bar{Y}$, the confidence interval is:

$$\bar{y} \pm ts(\bar{y}) = \bar{y} \pm t\frac{s}{\sqrt{n}}\sqrt{(\frac{N-n}{N})}$$

(Ignore the fpc if $n \leq 0.05N$)

The value t depends on the level of confidence desired. For large samples, the most common values (see Appendix I) are

t = 1.28 for 80% confidence level

t = 1.64 for 90% confidence level

t = 1.96 for 95% confidence level

t = 2.58 for 99% confidence level.

If the sample size is less than 30, the percentage points may be taken from the Student's t table (see Appendix II) with (n-1) degrees of freedom.

3.4.5 Approach to Normal Distribution

Comparing Tables 3.5 and 3.6, it can be seen that as the sample size increases, the sample estimates differ less and less from the expected value, and at the same time the standard error becomes smaller and smaller. In practical sampling problems, where a reasonably large sample is used (generally 30 or more cases), the distribution of sample results over all possible samples approximates very closely the **normal distribution**-- the familiar bell-shaped curve. This is the result of the most important theorem in statistics, The Central Limit Theorem, which states, briefly, that sums of random variables have a normal distribution.

For this distribution, the probabilities of being within a fixed range of the average value are well known and have been published (see Appendix I). These probabilities depend solely on the value of the standard error. For example, the probability of being within one standard error is 68 percent; for two standard errors, it is 95 percent; for three standard errors, it is 99.7 percent.

The implications are of fundamental importance to sampling theory. Suppose we have drawn a simple random sample from a population, have computed the mean from the sample $(\bar{y})$ and have estimated the true standard error of the mean $S(\bar{y})$, by means of $s(\bar{y})$. How can we infer the precision of this particular sample result? If we set an interval based on $s(\bar{y})$ around the sample estimate $(\bar{y})$, we can be fairly confident that $[\bar{y} \pm s(\bar{y})]$ will give an interval such that one
correct about two-thirds of the time that the interval covers the true mean.
Similarly, $[\bar{y} \pm 2s(\bar{y})]$ gives a confidence interval for which the assumption will be correct 95 percent of the time, and for $[\bar{y} \pm 3s(\bar{y})]$ it will be correct 99.7 percent of the time. To understand the concept, we present the following illustration.

3.4.6 Illustration

Consider again the same population of 12 individuals in Table 3.5. Let us find the percent of sample averages in Table 3.5 which differ from the population average $\bar{Y}$ = \$2,675 by less than $S(\bar{y}),\ 2S(\bar{y}),\ 3S(\bar{y})$. (We are using capital S instead of small s, as well as $\bar{Y}$, be
dealing with a population and we therefore know its true variance and its true mean). This is the same as finding the percent of sample averages which fall within $[\bar{Y} \pm S(\bar{y})],\ [\bar{Y} \pm 2S(\bar{y})],$ and $[\bar{Y} \pm 3S(\bar{y})]$. Consider a sample of size 2. Using Table 3.5 with $S(\bar{y})$ = \$1,015,

$$\bar{Y} - S(\bar{y}) = \$1,660 \qquad \bar{Y} + S(\bar{y}) = \$3,690$$

$$\bar{Y} - 2S(\bar{y}) = \$645 \qquad \bar{Y} + 2S(\bar{y}) = \$4,705$$

Table 3.6 shows that there are 42 sample averages that fall within the confidence interval (1660, 3690). That is, 64% of sample averages differ from the population average by less than one standard error. Similarly, there are 64 averages that fall within the confidence interval (645, 4705); that is, about 97% of sample averages differ from the population average by less than two standard errors. It can easily be seen that 100% of sample averages differ from the population average by less than three standard errors. For the normal distribution, we have seen that the probability of being within one standard error is 68%; for two standard errors, it is 95%; for three standard erros it is 99.7%. This shows that even for small samples of size 2, the distribution of sample results over all possible samples approximates very closely the normal distribution. For larger samples, the results would conform to the normal distribution much more closely. The percentages of sample averages in Table 3.5 which differ from the population averages by less than $S(\bar{y})$, $2S(\bar{y})$, $3S(\bar{y})$ and are displayed in Table 3.7.

Table 3.7

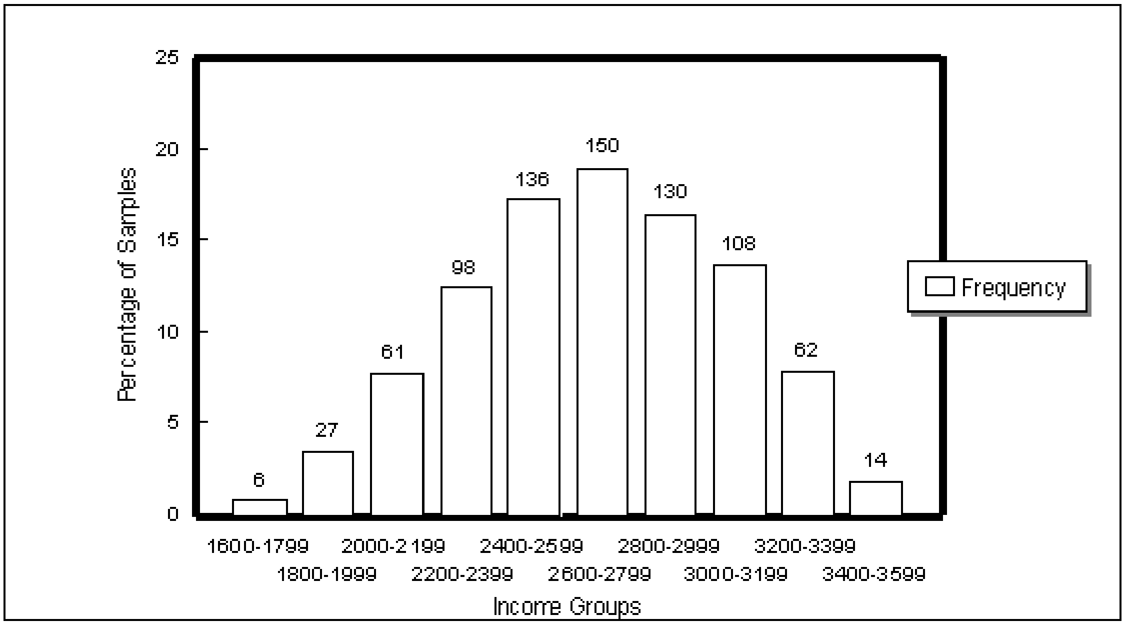
CONCENTRATION OF SAMPLE RESULTS AROUND THE POPULATION AVERAGE

Sample of Size n	$S(\bar{y})$	Percent of sample averages in Table 3.5 differing from the population average by		
		less than $S(\bar{y})$	less than $2S(\bar{y})$	less than $3S(\bar{y})$
1	$1,505	75	92	100
2	1,015	64	97	100
3	786	65	96	100
4	642	64	97	100
5	537	65	97	100
6	454	64	97	100
7	383	65	97	100
NORMAL DISTRIBUTION		68	95	99.7

Consider the distribution given in Table 3.5 of average income in all possible samples of size 7. A graph of this distribution is shown in Figure 3.1. This figure appears approximately symmetric, with a clustering of measurements about the midpoint of the distribution, tailing off rapidly as we move away from the center of the histogram. Thus, the graph possesses the following properties:

Figure 3.1

DISTRIBUTION OF AVERAGE INCOME IN ALL POSSIBLE SAMPLES OF SIZE 7



(1) The sampling distribution of $\bar{y}$ appears approximately normally distributed when the sample size is large.

(2) The average of all possible sample averages equals the population average.

(3) The variance of the sampling distribution is equal to $\left(\frac{S^2}{n}\right)\frac{(N-n)}{N}$ which is less than t population variance,

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

Property (1) above is the result of the Central Limit Theorem (CLT), one of the most fundamental and important theorems in statistics. Briefly stated, the CLT shows that if $x_1, x_2, \ldots, x_n$ are independent random variables having the same distribution with mean μ and variance σ^2, then for a large enough sample, the variable

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution (i.e., mean zero and variance one).

3.4.7 Illustration

Unoccupied seats on flights cause the airlines to lose revenue. Suppose a large airline wants to estimate the average number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected from the files, and the number of unoccupied seats is noted for each of the sampled flights.

The sample mean and standard deviation are

$$\bar{y} = 11.6 \text{ seats and } s = 4.1 \text{ seats}$$

Estimate $\bar{Y}$, the mean number of unoccupied seats per flight during the past year, using a 90% confidence interval (ignore the fpc).

The 90% confidence interval is,

$$\bar{y} \pm t\frac{s}{\sqrt{n}} = 11.6 \pm 1.645\frac{4.1}{\sqrt{225}} = 11.6 \pm .45 = (11.15,\ 12.05)$$

that is, at the 90% confidence level, we estimate the mean number of unoccupied seats per flight to be between 11.15 and 12.05 during the sampled year.

3.4.8 Sampling and Nonsampling Errors

Estimates are subject to both sampling errors and nonsampling errors. Sampling error arises because information is not collected from the entire target population, but rather from some portion of it. Through the use of scientific sampling procedures, however, it is possible to estimate from the sample data the range within which the true population value (parameter) is likely to be with a known probability.

Nonsampling error, on the other hand, is defined as a residual category consisting of all other errors which are not the result of the data having been collected from only a sample. These include errors made by respondents, enumerators, supervisors, office clerical staff, key coding operators, etc.

3.4.9 Total Error (Mean Square Error)

The total error is the sum of all errors about a sample estimate, both sampling and nonsampling, both variable and systematic. An illustration of the composition of the total error follows:

Total Error

Sampling Error Non Sampling Error

Variable Error Bias Variable Error **Bias**

In practice, the bulk of sampling error consists of variable error, and by contrast the bulk of nonsampling error is bias.

Mathematically, the total error is represented by the mean square error. In terms of expected values, the mean square error of the estimate $\hat{\theta}$ is denoted by the $MSE(\hat{\theta})$ and is given by:

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (Bias)^2$$

which is the average of the squares of deviations of all possible estimates from the parameter.

Recall that $Bias = [E(\hat{\theta}) - \theta].$ If the estimates are unbiased, the mean square error is to the variance.

STUDY ASSIGNMENT

Problem A: *You want to compute some averages (means) and standard deviations of the number of cows per farm. Assume that you know the number of cows per farm for each of eight farms, as follows:*

Farm	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Number of Cows	*4*	*5*	*0*	*3*	*2*	*1*	*1*	*0*

Exercise 1. *Calculate the average number of cows per farm.*

Exercise 2. *Calculate the standard deviation of the number of cows per farm.*

Exercise 3. *Take all possible samples of two farms each and calculate the average number of cows per farm for each sample.*

Exercise 4. *Prepare a frequency distribution showing the number of samples (of two farms each) for which the average for the sample falls in each of the following groups:*

Under 1.00
1.00 to 1.49
1.50 to 1.99
2.00 to 2.49
2.50 to 2.99
3.00 to 3.49
3.50 to 3.99
4.00 or more.

Exercise 5. *Compute the average of the 28 means obtained in exercise 3 and compare it with the true mean.*

Exercise 6. *Compute the standard error* $S(\bar{y})$ *for means of samples of 2 farms.*

Exercise 7. *Convert the data in Table 3.5 (on page 18) to percentage distributions for n = 1, 3, 5, and 7 (by dividing frequencies by the total shown on next to the last line of the table). Plot the histograms for n = 1, n = 3, n = 5, and n = 7 on the same chart (superimposed or parallel), using different colored pencils if necessary. Make proper adjustments for the fact that the first interval is twice as large as the standard interval, and the last interval is 8 times as large as the standard interval. Label the chart and the individual histograms.*

Notice how the distributions become more closely centered about the mean as n becomes larger, and also how the distributions approach the smooth normal distribution as n becomes larger.

Exercise 8. *Using exercises 5 and 6, find the proportions of the 28 values of* $\bar{y}$ *that are between*

$$[E(\bar{y})\pm S(\bar{y})],\ [E(\bar{y})\pm 2S(\bar{y})]\ \textit{and}\ [E(\bar{y})\pm 3S(\bar{y})].$$

the expected proportion assuming the sampling distribution of $\bar{y}$ *is normal?*

Problem B: *Consider the following distribution of N = 6 population values which represent "the number of household persons residing in the housing unit." Random samples of size 2 are drawn from this population.*

Housing Unit (HU)	*Household Size (HH) Size*
1	*5*
2	*6*
3	*7*
4	*8*
5	*9*
6	*10*

Exercise 9. *Show that the mean of this population is* $\bar{Y}$ *= 7.5 and its standard deviation is*

$$S = \sqrt{\frac{35}{12}}$$

Exercise 10. *How many possible random samples of size 2 can be drawn from this population? List them all and calculate their means.*

Exercise 11. *Use the results of exercise 10 to assign to each possible sample a probability and construct the sampling distribution of the mean for random samples of size 2 from the given population.*

Exercise 12. *Calculate the mean and the standard deviation of the probability distribution obtained in exercise 11.*

Exercise 13. *A simple random sample of 100 households will be selected from a village of Nigeria. For this village* $\bar{y}$ *= 75 Naira per month is spent on electricity and s = 15 Naira. Find a 95% confidence interval for* $\bar{Y}$. *Interpret the interval (ignore the fpc).*

Exercise 14. *A manufacturing company wishes to estimate the mean number of hours per month an employee is absent from work. The company decides to randomly sample 320 of its employees from a total of 5,000 employees and monitor their working time for 1 month. At the end of the month the total number of hours absent from work is recorded for each employee. If the mean and standard deviation of the sample are* $\bar{y}$ *= 9.6 hours and s = 6.4 hours, find a 95% confidence interval for the true mean number of hours absent per month per employee.*

CHAPTER 4

SIMPLE RANDOM SAMPLING
BASIC THEORY

4.1 SIMPLE RANDOM SAMPLING

The simplest method of probability sampling is simple random sampling (SRS).

To introduce the idea of a simple random sample, let us ask the following questions:

(1) How many distinct samples of size n can be drawn from a population of size N?

(2) How can we define a simple random sample?

(3) How can a random sample be drawn in actual practice?

To answer the first question, we use combinatorics, which allows us to choose n objects

out of a total of N ($n \le N$) $\binom{N}{n}$ in $= \dfrac{N!}{n!\ (N-n)!}$ ways, where N! = N (N-1) (N-2) ...

For instance, $\binom{5}{2} = \dfrac{5!}{2!\ 3!} = \dfrac{(5)(4)(3)(2)(1)}{(2)(1)\ (3)(2)(1)} = 10$ different s

drawn from a population of size N=5.

To answer the second question, we make use of the answer to the first one and define a <u>simple random sample</u> of size n (or more briefly, a <u>random sample</u>) selected from a population of size N

as a sample which is chosen in such a way that each of the $\binom{N}{n}$ possible samples has the same

probability of being selected. This probability is equal to:

$$\frac{1}{\binom{N}{n}}$$

For example, if a population consists of the N=5 elements A, B, C, D and E (which might be the incomes of five persons, the number of persons in five households, and so on), there

are $\binom{5}{3} = 10$ possible distinct samples of size n = 3; they consist of the elements ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, and CDE. If we choose one of these samples in such a way that each has the probability $\frac{1}{\binom{5}{3}} = \frac{1}{10}$ = 1/10 of being chosen, we call this sample a simple random sample.

With regard to the third question of how to take a random sample in actual practice, we could, in simple cases like the one above, write each of the $\binom{N}{n}$ possible samples on a slip of paper, put these slips into a hat, shuffle them thoroughly, and then draw one without looking. Such a procedure is obviously impractical, if not impossible, given the size of most populations; we mentioned it here only to make the point that the selection of a random sample must depend entirely on chance.

Fortunately, we can take a random sample without actually resorting to the tedious process of listing all possible samples. We can list instead the N individual elements of a population, and then take a random sample by choosing the elements to be included in the sample one at a time, making sure that in each of the successive drawings each of the remaining elements of the population has the same chance of being selected. The selection may be accomplished by either sampling with replacement or sampling without replacement. In sampling from a finite population, the practice usually is to sample without replacement. Most of the theory which will be discussed is based on this method. For example, to take a random sample of 12 of a city's 273 drugstores, we could write each store's name (address, or some other business identification number) on a slip of paper, put the slips of paper into a box or a bag and mix them thoroughly, and then draw (without looking) 12 of the slips one after the other without replacement.

Even this relatively easy procedure can be simplified in actual practice; usually, the simplest way to take a random sample from a population of N units is to refer to a table of random numbers (see Appendix III). In practice, however, the members of the population are sorted according to certain rules and then a systematic selection of n elements is carried out. The sample thus obtained is, for all practical purposes, a simple random sample.

4.1.1 Procedure for Selecting a Simple Random Sample (Use of Random Number Tables)

A practical procedure of selecting a random sample is to choose units one by one with the help of a table of random numbers. Tables of random numbers are used in practical sampling to avoid the necessity of carrying out some operation such as selecting numbered chips from an urn to designate the units to be included in the sample. Moreover, experience has shown that it is practically impossible to mix a set of chips thoroughly between each selection, that devices such as cards or dice have imperfections in their manufacture, that in thinking of numbers at random people tend to favor certain digits, etc. Consequently, such methods do not, in fact, give each member of the population an equal chance of selection. The use of a table of random numbers, however, reduces the amount of work involved, and also gives much greater assurance that all elements have the same probability of selection.

Many tables of random numbers are readily available. There are several in the series of Tracts for Computers, notably tables compiled by Tippett, and by Kendall and Smith. The RAND Corporation has published A Million Random Digits. Sets are also available in Statistical Tables by Fisher and Yates, and in other sources. Many of these publications describe the methods of compilation and the uses of the tables. Some microcomputer packages such as LOTUS spreadsheets also have a random number generator which can also be used to generate pseudo-random numbers, but these random number generators provide random numbers between 0 and 1. A table of random numbers is given in Appendix III.

Typically, these tables show sets of random digits arranged in groups both horizontally and vertically. To select a set of random numbers, one can start anywhere on a page. Furthermore, after selecting the first number, one can proceed down a column, across a row, up a column, or in any other pattern that is desired.

4.1.2 Illustration

To obtain a random number between 1 and a given number, for example between 1 and 273, proceed as follows: Notice how many digits are in the upper limit number (for 273 there are three digits). Use this number of columns, counting from the first (or a predetermined) column, and start at the top (or on a predetermined line). Each line in the set of three columns has a 3-digit number. Choose the first of these which is between 001 and the given number, inclusive. That is, between 001 and 273 in our example. Discard numbers which are greater than 273 and discard 000. If more than one random number is desired, continue down the three columns, choosing each 3-digit number which is between 001 and 273 until the desired 3-digit random number is obtained. If a number is chosen two or more times, use it only once.[3]

Suppose we have a part of a table of random numbers as follows:

1 0 8 9	8 7 1 9
9 3 8 5	7 9 0 2
6 9 3 4	8 6 6 0
0 0 5 2	1 0 0 7
5 7 3 6	9 2 4 9
1 9 0 1	5 9 8 8
5 3 7 2	6 2 1 2

Within the limits of the numbers in the examples which follow, we shall select random numbers from the above table, using a selected number only once.

Example A: Select 3 numbers at random between 1 and 10. First choose an arbitrary column, having decided to let 0 stand for 10. Suppose we choose the fifth column. The first number in the

[3] There is a variation of this method that saves time when the upper limit number is a power of 10, for example 10, 100, etc. In this case we can use 1 less column than the digits in the upper limit number; that is, use 1 column instead of 2 for choosing random numbers from 1 to 10; if a zero occurs, use it as the upper limit number. Thus in selecting random numbers from 1 to 10, use one column and take every number until the desired number of random numbers is obtained; if 0 occurs, treat it as if it were 10. For 1 to 100, use two columns and treat 00 as 100.

column is 8; the second number is 7; the third is 8 again. Since 8 has already been selected, we skip it and take the next number which is 1. The three numbers selected, therefore, are 8, 7, and 1.

Example B: Select 5 numbers at random between 1 and 80. Suppose we take the first two columns as our choice of a start. First take 10; discard 93 since it is not between 01 and 80; take 69; discard 00 (which represents 100); and take 57, 19, and 53.

4.1.3 Caution in the Use of Random Table

If we use a table of random numbers frequently, we should not always use the same part. For example, if the first random number is always taken from the same column of the same page, the same set of numbers would be used repeatedly, and we would not get proper randomization. If tables of random numbers are used frequently, one can continue from the last random number selected for the previous sample, or a new starting point should be taken for each use.

4.2 NOTATION

The notation defined in this section is appropriate not only for simple random sampling, but also for most designs. They provide a key to the system used throughout this manual. Capital letters refer to population values and lower case (small) letters denote corresponding sample values. A bar (-) over a letter denotes an average or mean value and (^) over a letter indicates an estimate. We shall use the following notation:

N = total number of units in the population

n = total number of units in the sample

Y_i = value of a characteristic as measured on the i-th unit in the population; i=1, 2, ... N

y_i = value of a characteristic as measured on the i-th sample unit; i = 1, 2,...n

$Y = \sum_{i=1}^{N} Y_i$ total value of a characteristic in the population

$y = \sum_{i=1}^{n} y_i$ total value of a characteristic in the sample (or sum of sample values for the characteristic)

$\bar{Y} = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ = population mean

$\bar{y} = \frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$ = sample mean

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ = population variance

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ = population variance

$S^2 = \frac{N}{N-1}\sigma^2 \qquad \sigma^2 = \frac{N-1}{N}S2$, and

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2) =$ sample

$f = \frac{n}{N}$ = sampling rate or sampling fraction

$\frac{1}{f} = \frac{N}{n}$ sampling weight (expansion factor)

CV = coefficient of variation

cv = estimated coefficient of variation

As we mentioned earlier, we shall use, unless otherwise mentioned, S^2 for the population variance. The difference between S^2 and σ^2 disappears for large populations. In general, the population variance, S^2, is not known. The sample variance, s^2, will be used as its estimate; this will hold throughout the course regardless of the sampling scheme being discussed. It should be noted that in simple random sampling, s^2, is an unbiased estimate of S^2.

4.2.1 Population Values, Their Respective Estimates, and Measures of Precision

The sample estimate of the population total value, Y, is denoted by $\hat{Y}$, and can be written as:

(4.1) $$\hat{Y} = N\bar{y} = \frac{N}{n}y = \frac{N}{n}\sum_{i=1}^{n} y_i$$

where $\bar{y}$ is the estimate of the population average, $\bar{Y}$, and is given by:

(4.2) $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The standard error of the estimate of $\hat{Y}$ is:

(4.3) $$S(\hat{Y}) = \frac{NS}{\sqrt{n}}\sqrt{(\frac{N-n}{N})}$$

and the standard error for$\bar{y}$ is:

(4.4) $$S(\bar{y}) = \frac{S}{\sqrt{n}}\sqrt{(\frac{N-n}{N})}$$

The corresponding formulas for the estimated standard error are:

(4.5) $$s(\hat{Y}) = \frac{Ns}{\sqrt{n}}\sqrt{(\frac{N-n}{N})}$$

(4.6) $$s(\bar{y}) = \frac{s}{\sqrt{n}}\sqrt{(\frac{N-n}{N})}$$

4.2.2 Illustration

Let us verify equation (4.3) with the data for the 12 individuals discussed previously (see Chapter 3). We have already used equation (4.4) for the means of samples of sizes 1 and 2 in illustration 4.3, and their standard errors for different sizes were given in Table 3.6 of Chapter 3. Using this table, the total income of 12 individuals can be estimated. Equation (4.3) can be expressed as:

$$S(\hat{Y}) = \frac{NS}{\sqrt{n}}\sqrt{(\frac{N-n}{N})} = NS(\bar{y})$$

Using Table 3.6 of Chapter 3, the standard error of the estimated total income for samples of size 2 is:

$$S(\hat{Y}) = 12 \ x \ (\$1{,}015) = \$12{,}180..$$

4.2.3 Relative Error

Often we wish to consider not the absolute value of the standard error, but its value in relation to the magnitude of the statistic (mean, total, etc.) being estimated. For this purpose, one can express the standard error as a proportion (or a percent) of the value being estimated. This form is called the relative standard error, or coefficient of variation and is denoted by the symbol CV, with parentheses to indicate the statistic to which the error applies. One advantage of expressing error as CV's is that it is unitless, unlike absolute measures like the standard deviation and the standard error. The CV is useful when making comparisons because no units enter into play. The population CV refers to the relative standard error of means of samples of 1 unit (that is, the population standard deviation expressed as a proportion of the population mean) and it's denoted simple by CV (not followed by a parenthesis). Thus, for the estimate of the total, the true[4] coefficient of variation is:

$$CV(\hat{Y}) = \frac{S(\hat{Y})}{\hat{Y}} = \frac{\frac{NS}{\sqrt{n}}\sqrt{\frac{N-n}{N}}}{N\bar{y}} \tag{4.7}$$

$$= \frac{1}{\sqrt{n}}\frac{S}{\bar{y}}\sqrt{\frac{(N-n)}{N}} \quad \frac{CV}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}} =$$

Similarly, for the estimate of the sample mean, the coefficient of variation is:

$$CV(\bar{y}) = \frac{S(\bar{y})}{\bar{y}} = \frac{1}{\sqrt{n}}\frac{S\sqrt{\frac{N-n}{N}}}{\bar{y}} \tag{4.8}$$

$$= \frac{CV}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}}$$

Notice that $CV(\hat{Y}) = CV(\bar{y})$

The corresponding formulas for the estimated coefficient of variations are:

$$cv(\hat{Y}) = \frac{s(\bar{y})}{\bar{y}} = \frac{Ns}{N\bar{y}\sqrt{n}}\sqrt{\frac{N-n}{N}} = \frac{1}{\sqrt{n}}\frac{s}{\bar{y}}\sqrt{\frac{N-n}{N}} \tag{4.9}$$

$$cv(\bar{y}) = \frac{s(\bar{y})}{\bar{y}} = \frac{s}{\bar{y}\sqrt{n}}\sqrt{(\frac{N-n}{N})} \tag{4.10}$$

The standard error of the estimated total is N times that for the mean, while the coefficients of

[4] Corresponding estimates from the sample are obtained by substituting s^2 for S^2 $\bar{y}$,$\bar{Y}$ for $\hat{Y}$ and for Y in equations (4.7) and (4.8).

variation of the two estimates are the same; this result is, upon reflection, not unexpected. An estimated total is obtained by multiplying the sample mean (an estimate) by the number of elements in the population (a known number); the only source of error is the sample mean. Therefore, we should expect that, when expressed as a proportion or percentage, the error in the total would be the same as that in the mean; however, when the error in the total is expressed in absolute terms, it would be N times as large as the error in the mean, since N is the factor of multiplication.

The big advantage of the coefficient of variation is that it permits comparison of two distributions of values even though they may be totally unrelated. For example, one could compare the variability of length of mice tails to weight of elephants. This is possible because variability is expressed relative to the mean, that is, it is the average variability per unit of mean.

Another way to look at the cv is to consider it as a measure of dispersion for relative deviations. Recall that the variance of Y_i was given by

$$\sigma^2 = \frac{\sum (Y_i - \bar{Y})^2}{N}$$

This is a measure of dispersion of the absolute deviations $(Y_i - \bar{Y})$.

If we now consider the relative deviations

$$\frac{Y_i - \bar{Y}}{\bar{Y}}$$

square them, add them, and then average them over N, we get the following expression:

$$(CV)^2 = \frac{\sum_{i=1}^{N} \left(\frac{Y_i - \bar{Y}}{\bar{Y}} \right)^2}{N}$$

which is called the ***relative variance*** of the distribution or simply the ***relvariance***.

If we rearrange terms in the above expression, we get:

$$(CV)^2 = \frac{\dfrac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}{N}}{(\bar{Y})^2} = \frac{\sigma^2}{\bar{Y}^2}$$

The square root of this last expression is the population coefficient of variation mentioned before.

4.3 SAMPLING FOR PROPORTIONS

An important class of statistics for which the formulas for variance and the formulas for determining the size of sample become particularly simple is the estimation of the proportion of units having a certain characteristic.

4.3.1 Types of Statistics for Which Proportions are Used

Proportions arise in two ways in statistical analysis. First of all, we are frequently interested in a statistic that is a proportion, rather than a total or an average; for example, the proportion of the population that is unemployed, or the percentage of families with income greater than a certain amount, or the proportion of business firms interested in purchasing a particular product. Secondly, it may be desired to classify a population into a number of groups, and to find the percentage of the total population in each of these groups. The groups may have a natural ordering as in distribution by age (0 to 4 years, 5 to 9, 10 to 14, etc.) or income classes; or they may be groups having no natural order, such as those in an industrial classification of business firms, where the groups can be arranged in a number of ways. The analysis is the same whenever the proportion of the total in each group is the statistic to be measured.

4.3.2 Relationship to Previous Theory

Suppose we think of the total population and the sample in the following way. Consider a particular class of units in which we are interested, and use the following notation:

A = Total number of units in that class in the population
a = Number of units in that class in the sample

P = True proportion of units in that class in the population
p = Proportion in that class in the sample

Q = Population proportion not in that class (Q = 1 - P)
q = Proportion not in that class in the sample (q = 1 - p).

Note that P = $\frac{A}{N}$ and p$\frac{a}{n}$.

All of the formulas discussed in previous lectures can be applied to this particular case by considering each member of the population as having a characteristic which can have only one of two values, either 0 or 1. If the member is in a particular class in which we are interested, the value assigned is 1; if the member is not in the class, the value is 0. Examining the entire population, we can see that the A members of the class each have a value of 1; the rest have a value of 0. Adding up the values for all elements of the population, we get A. In other words, A can be considered as the equivalent of

Y = $\sum_{i=1}^{N} Y_i$ that we have discussed. Similarly P = $\frac{A}{N}$ can be considered in the same way as $\bar{Y} = \frac{Y}{N}$. We can now use the previous formulas. It turns out that they are particularly easy to use in this case.

4.3.3 Applicable Formulas

In sampling for proportions, the following formulas are applicable (with simple random sampling):

(4.11) $\hat{P} = p = \frac{a}{n}$ and $\hat{A} = pN$

That is, an estimate of the proportion in the population is obtained by using the sample proportion, and an estimate of the total number of units having the characteristic is obtained by multiplying the sample proportion by the total number of units in the population. Also

(4.12) $$\sigma^2 = PQ; \quad S^2 = \frac{NPQ}{N-1}$$

The population variance is PQ. Note that it is the variance of the population distribution giving the value of 1 or 0 to an element depending on whether or not it is in the class (whether it has the attribute in question). It can still be estimated by pq, unless n is very small (for example n < 30) in which case the formula is s^2 = pq($\frac{n}{n-1}$).

The variance of the estimate of the proportion which is computed from all samples of size n is

(4.13) $$\sigma^2(\hat{P}) = \left(\frac{N-n}{N-1}\right)\frac{PQ}{n}$$

The estimate of this variance which is made from a single sample of n observations is

(4.13a) $$s^2(\hat{P}) = \frac{pq}{n-1}\left(\frac{N-n}{N}\right)$$

See equations (4.4) and (4.6). These are the same formulas given previously for $S^2(\bar{y}\ )$, with PQ substituted for S^2, and $s^2(\bar{y}\ )$, with pq substituted for s^2.

Similarly, the formulas given in the previous section for the relative standard error (coefficient of variation) of a mean and the standard error of an estimated total are given by:

$$(4.14) \qquad CV(\hat{P}) = \sqrt{\frac{Q}{Pn}\left(\frac{N-n}{N-1}\right)}, \quad cv(p) = \sqrt{\frac{q}{p(n-1)}\left(\frac{N-n}{N}\right)}$$

and,

$$(4.15) \qquad S(\hat{A}) = N\sqrt{\frac{PQ}{n}\left(\frac{N-n}{N-1}\right)}, \quad s(\hat{A}) = N\sqrt{\frac{pq}{(n-1)}\left(\frac{N-n}{N}\right)}$$

Again the relative standard error of the total is the same as that of the mean.

The confidence interval for the proportion is derived on the same assumptions as for the quantitative characteristics, namely that the sample proportion p is normally distributed. From (4.13a) for the estimated variance of p, one form of the normal approximation to the confidence interval for p is:

$$(4.16) \qquad \left(p \pm t\sqrt{\frac{pq}{n}\left(\frac{N-n}{N}\right)}\right)$$

where the value t depends on the level of confidence desired (see Section 4.4 of Chapter 3).

4.3.4 Illustration

Estimate of sampling error.--Suppose that the proportion of farms that grow maize in a given area is .40; what would be the sampling error in estimating this proportion from a random sample of 500 farms, if the total number of farms in the area is 10,000? In this case,

N = 10,000	P = 0.40
n = 500	Q = 0.60

We have

$$S^2(\hat{P}) = \frac{PQ}{n}\left(\frac{N-n}{N}\right) = \left(\frac{(.4)\ (.6)}{500}\right)\left(\frac{10{,}000-500}{10{,}000}\right) = \left(.\frac{24}{500}\right)\left(\frac{9{,}500}{10{,}000}\right) = 0.000456$$

Consequently, $S(\hat{P}) = \sqrt{.000456} = .021$

How is the figure of .021 to be interpreted? This means that if we establish an interval around the true proportion of [.40 ± .0 21] (or .379 to .421), there is a reasonably good chance (68 percent) that a sample of 500 farms will give a proportion somewhere between .379 and .421. If we double the interval to get a range of .358 to .442, the chance is about 95 percent that the sample estimate will be within that range. If an interval based on three times .021, (or .063) is used, the chance is .997 (or nearly certain) that the sample estimate will be within that range. In normal practice, it is customary to use a 2-S range (two standard errors) as providing sufficient confidence in the accuracy of the estimates. If very important decisions are to be based on the results of the survey, and we wish to be almost absolutely sure of the range within which the sample estimate will lie, we can use a 3-S level. It is difficult to conceive of cases in which 3-S would not be sufficient.

In this example, both the proportion (.40) and the chance of the sample estimate being within a certain range around this proportion were known. In practice, we are usually interested in the converse of this situation, in which we do not know the true proportion but we do have a sample estimate of .40 based on a sample of 500 farms out of 10,000. We wish to establish ranges around the sample figure which will be expected to include the true mean. For all practical purposes, the same statements can be made as before by substituting the term "true figure" for "sample estimate." That is, if the sample shows that .40 of the farms grow corn and we establish a range [.40 ± .021] , the chances are about 68 percent that this range will include the true figure; the chances are about 95 percent that the interval .358 to .442 will include the true figure; etc.

4.3.5 Procedure When P Refers to a Subset of a Class

Frequently, the proportion to be estimated is a percentage, not of the total population, but of a particular class. For example, we may be interested not in unemployment expressed as a percentage of the total population, but as a percentage of persons in the labor force; or we may need to know the proportion of firms with more than 5 employees in a particular industry. In such cases, a very close approximation to an exact analysis can be made by using the formulas listed above, but interpreting the numbers N and n as applying to the class in which we are interested. That is, N would not be considered the total population but would be the number of persons in this class (for example, the total number of persons in the labor force) as estimated from the sample; n would be the number of sample cases in this class; a would be the number of sample cases in the subset (for example, the number of unemployed).

4.3.6 Tabled Value of $\sqrt{\frac{PQ}{n}}$

Table 4.1 shows the value of $\sqrt{\frac{PQ}{n}}$ for specified values of P and n. As described in sections 4.3.7 and 4.3.8 below, we can use the simplified formula

$$(4.17) \qquad S(\hat{P}) = \sqrt{\frac{PQ}{n}}$$

to compute the standard error of the proportion of units having a certain attribute, if the sample is an unrestricted (simple) random sample and if N is so large relative to n that the factor (N-n)/N in the formula has a value very close to 1.

Since the true proportion in the population (P) is not known, the estimate from the sample (p) may be used in equation (4.17) to give an estimate of the standard error of p:

$$(4.18) \qquad s(\hat{P}) = \sqrt{\frac{pq}{n-1}}$$

Most samples are stratified; that is, they are not simple random samples. We shall see in chapters 7 and 8 that this has the effect of making the sampling error smaller than it would be for a simple random sample of the same size. However, most samples used in surveys are also clustered and we shall see in chapters 9 and 10 that this has the opposite effect of making the sampling error larger than it would be for a simple random sample of the same size. When the sample is both stratified and clustered, the formulas for the standard error become more complex (see later chapters).

Sometimes it is not possible to work out the exact formulas, but a rough estimate of the standard error can be obtained by using the simple formula of equation (4.17) with an allowance for the expected net effect of departures from randomness in the sample design. If the units of analysis are clustered into rather small groups--for example, 5 housing units or 25 persons in a cluster, and the persons within a cluster are rather similar, as in a cluster located in a rural area--the standard error of a proportion as read from Table 4.1 might be multiplied by a factor such as 1.25. This factor is a design effect (see Section 3.6.) In a larger cluster, such as a city block with 40 or 50 housing units, the factor to be applied to Table 4.1 might be 1.75, even though the persons within the cluster are less alike in an urban area than in a rural area.

The size of the design effect to be used depends on the sample design and the nature of the population; it can sometimes be roughly estimated by an experienced sampling statistician, using past experience and mathematical formulas involving the "intraclass correlation" (see Chapter 10).

Table 4.1
STANDARD ERROR OF AN ESTIMATE OF A PROPORTION IN SIMPLE RANDOM SAMPLING

$(S(\hat{P})$ $\sqrt{\frac{PQ}{n}}$ for specified values of P and n)

P = Proportion of units[5] having a characteristic (Q = 1-P has the same standard error)													
n=number of sample cases[6]	.001 or .999	.002 or .998	.01 or .99	.02 or .98	.03 or .97	.04 or .96	.05 or .95	.10 or .90	.15 or .85	.20 or .80	.25 or .75	.30 or .70	.40 or .60	.50
50	.0045	.0063	.0141	.0198	.024	.028	.031	.042	.051	.057	.061	.065	.069	.071
100	.0032	.0045	.0099	.0140	.017	.020	.022	.030	.036	.040	.043	.046	.049	.05.0
200	.0022	.0032	.0071	.0099	.012	.014	.016	.021	.025	.028	.031	.033	.035	.035
300	.0018	.0026	.0058	.0081	.0099	.012	.013	.017	.021	.023	.025	.027	.028	.029
400	.0016	.0023	.0050	.0070	.0086	.010	.011	.015	.018	.020	.022	.023	.024	.025
500	.0014	.0020	.0045	.0063	.0076	.0089	.0098	.013	.016	.018	.019	.021	.022	.022
600	.0013	.0018	.0041	.0057	.0070	.0082	.0090	.012	.015	.016	.018	.019	.020	.020
700	.0012	.0017	.0038	.0053	.0065	.0076	.0083	.011	.014	.015	.016	.017	.019	.019
800	.0011	.0016	.0035	..0050	.0061	.0071	.0078	.011	.013	.014	.015	.016	.017	.018
1000	.0010	.0014	.0032	.0044	.0054	.0063	.0070	.0095	.011	.013	.014	.015	.015	.016
1200	.0009	.0013	.0029	.0040	.0049	.0058	.0064	.0087	.010	.012	.013	.013	.014	.014
1500	.0008	.0012	.0026	.0036	.0044	.0052	.0057	.0077	.0093	.010	.011	.012	.013	.013
1700	.0008	.0011	.0024	.0034	.0042	.0049	.0053	.0073	.0087	.0097	.011	.011	.012	.012
2000	.0007	.0010	.0022	.0031	.0038	.0045	.0049	.0067	.0081	.0090	.0097	.010	.011	.011
2500	.0006	.0009	.0020	.0028	.0034	.0040	.0044	.0060	.0072	.0080	.0087	.0092	.0098	.0100
3000	.0006	.0008	.0018	.0026	.0031	.0039	.0040	.0055	.0066	.0073	.0079	.0084	.0090	.0092
3500	.0005	.0008	.0017	.0024	.0029	.0034	.0037	.0051	.0061	.0068	.0073	.0078	.0083	.0084
4000	.0005	.0007	.0016	.0022	.0027	.0032	.0035	.0047	.0057	.0063	.0068	.0073	.0077	.0079
4500	.0005	.0006	.0015	.0021	.0025	.0030	.0033	.0045	.0054	.0060	.0065	.0069	.0073	.0074
5000	.0004	.0006	.0014	.0020	.0024	.0028	.0031	.0042	.0051	.0057	.0061	.0065	.0069	.0071

[3] In practice the sample value p would be used, inasmuch as the population value P would not be known.

[4] Values of n greater than 5,000: When n is multiplied by 100, the standard error is divided by 10.

4.3.7 The Design Effect (DEFF)

The design effect or DEFF is the ratio of the variance of the estimate obtained from the more complex sample (described later in this text) to the variance of the estimate obtained from a simple random sample of the same size. For instance, if $Var_D(\bar{y})$ is the variance of the estimate, say $\bar{y}$, obtained from a complex sample, and $\frac{N-n}{N}\frac{S^2}{n}$ is the variance of the same estimate based on a simple random sampling, then

$$DEFF = \frac{Var_D(\bar{y})}{\frac{(N-n)}{N}\frac{S^2}{n}} \qquad\qquad Var_D(\bar{y}) = (\frac{N-n}{N}\,\text{and}\,\frac{S^2}{n})(DEFF)$$

where Var_D = variance obtained from the more complex design

This approach is commonly used by practical samplers. For many situations where we can not estimate directly the variance of the estimate, we may be able to guess fairly well both the element variance S^2 and DEFF from experience with similar past data. This comprehensive factor attempts to summarize the effects of various complexities in the sample design especially those of clustering (see Chapters 9 and 10).

4.3.8 Finite Correction Factor (or Finite Population Correction Factor)

The exact formula for the relative variance (square of the coefficient of variation) of a mean for a simple random sample,

$$CV^2(\bar{y}) = (\frac{N-n}{N-1})\frac{CV^2}{n}$$

or

$$CV^2(\hat{P}) = (\frac{N-n}{N-1})\frac{Q}{Pn}$$

can be divided into two parts:

$$\frac{N-n}{N-1} \qquad \text{and}\ \frac{CV^2}{n} \qquad \frac{Q}{Pn} \qquad \text{or}$$

The only way the size of the total population comes into the formula is in the expression

$$\frac{N-n}{N-1}$$

This is usually called the finite population correction factor (fpc). If the population were infinite this factor would be 1 and the formulas would be much simpler:

$$CV^2(\bar{y}) = \frac{CV^2}{n} \qquad CV^2(\hat{P}) = \frac{Q}{Pn} \text{ or}$$

The value of $\frac{N-n}{N-1}$ is approximately equal to $1 - \frac{n}{N} = 1 - f$, *where* $\frac{n}{N} = f$

If the sampling rate is small, say less than .05, the effect of the finite population correction factor is very small and, for all practical purposes, the finite population correction factor can be ignored.

4.3.9 Simplification for Large Populations

With large populations and small sampling rates, the fpc can be ignored and the formulas become simpler.

Simplified Formula

	True Value	Estimate
Variance of the mean	$S^2(\bar{y}) = \frac{S^2}{n}$	$s^2(\bar{y}) = \frac{s^2}{n}$
Variance of a proportion	$S^2(\hat{P}) = \frac{PQ}{n}$	$s^2(\hat{P}) = \frac{pq}{n-1}$
Coefficient of variation of the mean	$CV^2(\bar{y}) = \frac{CV^2}{n}$	$cv^2(\bar{y}) = \frac{cv^2}{n-1}$
Coefficient of variation of a proportion	$CV^2(\hat{P}) = \frac{Q}{Pn}$	$cv^2(\hat{P}) = \frac{q}{p(n-1)}$
Variance of a total	$S^2(\hat{Y}) = N^2\frac{S^2}{n}$	$s^2(\hat{Y}) = N^2\frac{s^2}{n}$
Variance of the total number of units having an attribute	$S^2(\hat{A}) = N^2\frac{PQ}{n}$	$s^2(\hat{A}) = N^2\frac{pq}{n-1}$
Coefficient of variation of a total	$CV^2(\hat{Y}) = CV^2(\bar{y}) = \frac{CV^2}{n}$	$cv^2(\hat{Y}) = cv^2(\bar{y}) = \frac{cv^2}{n}$
Coefficient of variation of total number of units having an attribute	$CV^2(\hat{A}) = CV^2(\hat{P}) = \frac{Q}{Pn}$	$cv^2(\hat{A}) = cv^2(\hat{P}) = \frac{q}{p(n-1)}$

Study Assignment

Problem A: *Explain why each of the following samples does not qualify as a random sample from the required population or might otherwise fail to give the desired information:*

Exercise 1. *To predict a municipal election, a public opinion poll telephones persons selected haphazardly from the city's telephone directory.*

Exercise 2. *To determine public sentiment about certain foreign trade agreements, an interviewer asks voters: "Do you feel that this unfair practice should be stopped?*

Exercise 3. *To determine the average annual income of its graduates 10 years after graduation, a university's alumni office sent questionnaires in 1992 to all members of the class of 1982, and the estimate is based on the questionnaires returned.*

Exercise 4. *To study consumer reaction to a new convenience food store, a house-to-house survey is conducted during weekday mornings, with no provisions for return visits in case no one is at home.*

Exercise 5. *How many different samples of size $n = 3$ can be selected from a population of size $N = 10$? $N = 25$? $N = 50$?*

Problem B: *Given a population of 185 persons, you want to select a sample from this population.*

Exercise 6. *Select a simple random sample of 20 persons. Use columns 1, 2, and 3 of the table of random numbers; use columns 4, 5, and 6 if columns 1, 2, and 3 are not sufficient; continue using successive columns as needed. List the numbers assigned to the 20 persons selected and describe the procedure you used in the selection.*

Problem C: *Refer to the data for the eight farms given in the problem for chapter 3.*

Exercise 7. *Suppose you have information on the number of cows from a sample of three farms--Farm No. 1, Farm No. 2, and Farm No. 8. You also know that there are eight farms in the group (population) from which this sample was drawn. Estimate the number of cows on all eight farms.*

Exercise 8. *Using the formula for the standard error of the mean, compute the standard error of the average number of cows per farm for a sample of two farms. (Use equation 4.4 since N is not large.)*

Exercise 9. *Using the frequency distribution developed in exercise 4 of chapter 3, determine*

(a) What proportion of samples (of two farms each) is within plus or minus one standard error of the average?

(b) What proportion is within plus or minus two standard errors?

(c) What are the proportions you would expect on the basis of the theory?

Problem D: *A simple random sample of employed persons is selected at a rate of one percent. The number of employed persons in the sample is 30,000; of these, 12,000 are employed in agriculture.*

Exercise 10. *Of the employed persons, what proportion is engaged in agriculture and what is the sampling error of this proportion?*

Problem E: *The total population of a country is 10,000,000. A 1/10 of one percent simple random sample is selected, amounting to 10,000 persons. In the sample, 4,000 persons are in the labor force, of whom 200 are unemployed.*

Exercise 11. *What proportion of the labor force in the country is unemployed and what is the standard error of this proportion?*

Problem F: *The total population of a city is 50,000. A 20-percent simple random sample is selected, amounting to 10,000 persons. In the sample, 4,000 persons are in the labor force, of whom 200 are unemployed.*

Exercise 12. *What proportion of the labor force in the city is unemployed and what is the standard error of this proportion?*

Exercise 13. *Comment on the reason for the difference in the standard errors in exercises 10 and 11.*

Exercise 14. *In a simple random sample of 200 from a population of 2,000 colleges, 120 colleges were in favor of a proposal, 57 were opposed, and 23 had no opinion. Estimate the 95% confidence interval for the number of colleges in the population that favored the proposal. Do the results of the sample furnish conclusive evidence that the majority of the colleges in the population favored this proposal?*

Problem G: *The following table shows the distribution of the households by region, from the 1988 Senegal Census frame*[7]*, with the corresponding number of sample households and number of economically active persons in the sample. An unemployment rate (p) of 20 percent with a design effect (DEFF)*[8] *of 2.5 is assumed for all regions:*

Region	Total no. HHs	Total no. of Sample HHs	Number of economically active persons in sample
Dakar	193,968	2,240	5,543
Thies	97,962	840	2,814
Louga	52,559	400	1,303
Kaolark	83,775	650	2,245
Fatick	55,041	410	1,322
Diourbel	66,213	510	1,664
Ziguinchor	53,489	450	1,172
Kolda	60,121	460	1,547
St. Louis	77,493	650	1,911
Tambacounda	42,998	410	1,280

Exercise 15. *Compute the overall sampling rates of sample households by region.*

[3] Source: Preliminary Recomendations for Designing the Master Sample Frame for the Senegal Intercensal Household Survey Program; David J. Megill, U.S. Bureau of the Census, November 1990.

[4] The design effect (see section 3.7) measures the effect of the stratification and clustering in the sample, compared to a simple random sample. The design effect is used here because a multistage stratified sample was used.

Exercise 16. *Using the formula* $s(\hat{p}) = \sqrt{(\frac{pq}{n}) \times DEFF}$, *where n is the number of economically persons in the sample, compute the standard error for the estimate of the unemployment rate for each region.*

Exercise 17. *Compute the coefficient of variation for the estimate of the unemployment rate for each region.*

Exercise 18. *Find the 95% confidence interval for the estimate of the unemployment rate for each region.*

Problem H: *Consider Appendix IV containing a list of 600 households residing in 30 villages located in 3 zones. The data on the total population and the size of the household were obtained during a census conducted 5 years ago. We have the following population values:*

Total number of zones = 3
Total number of villages = 30
Total number of households = 600
Total number of persons = 3037
Total number of persons from previous census = 2815
Total area = 270 km²
Average number of households per village = 20
Average number of persons per village = 101.42
Average number of persons per household = 5.07.

Select a simple random sample of 20 households without replacement, and on the basis of the data on the size of these 20 sample households, do the following for the entire population:

Exercise 19. *Estimate the total number of persons and its standard error.*

Exercise 20. *Compute the average household size (mean number of persons per household) and its standard error.*

Exercise 21. *Compute the coefficient of variation for each of the above.*

Exercise 22. *Find the 95% confidence interval for each of the above.*

CHAPTER 5

SIMPLE RANDOM SAMPLING ESTIMATION OF SAMPLE SIZE

5.1 SPECIFIC CONSIDERATIONS FOR DETERMINING THE SAMPLE SIZE

One of the first questions which a statistician is called upon to answer in planning a sample survey refers to the size of the sample required for estimating a population parameter with a specified precision. Making a decision about the size of the sample for the survey is important. Too large a sample implies a waste of resources, and too small a sample diminishes the utility of the results.

When considering sample size determination, there are three very important concerns: ACCURACY, PRACTICALITY, and EFFICIENCY.

5.1.1. Accuracy

Accuracy can be defined as the inverse of the total error. Total error is the sum of sampling error (SE) and nonsampling error (NSE). Sampling error arises because only a part of the population is observed, and not all of it. The terms PRECISION and RELIABILITY are associated with sampling error. Estimator A is more precise or more reliable than estimator B if the sampling error of A is smaller than the sampling error of B. Nonsampling errors are usually biases which are very often due to poor quality control of the survey operations (poor questionnaire; interviewers that are not well trained; response errors; etc.)

5.1.2. Practicality

To obtain an accurate estimate, both sampling and nonsampling errors must be reduced. However, accuracy may come into conflict with practicality because:

a. to reduce sampling errors and increase precision, the sample size must be large.

b. too large a sample can impose an excessive burden on the limited resources available (and resources are usually very limited) and increase the likelihood of nonsampling errors.

5.1.3. Efficiency

A further concern is that a given sample size can produce different levels of precision depending on which sampling techniques are chosen. This concept is known as the statistical efficiency of the design. The most efficient design is the one that gives the most precision for the same sample size. Therefore, expert sample design is needed in the determination of the optimal sample size.

Example 5.1

A population consists of N = 5000 persons. A simple random sample without replacement (SRS-WOR) of size n = 50 included 10 persons of Chinese descent.

A 95% confidence interval for P, the proportion of persons of Chinese descent in the population, is:

$$p \pm 2\sqrt{(1-f)\frac{pq}{n}}$$

$$0.20 \pm 2\sqrt{\left(1 - \frac{50}{5000}\right)\ \frac{(.20)(.80)}{50}} = (0.087,\ 0.312)$$

The conclusion is that between 8.7% and 31.2% of the population is of Chinese descent. This interval is too wide to be useful. There are two ways in which a narrower interval could be obtained:

- by lowering the confidence level, or
- by increasing the sample size

There is a point at which lowering the confidence level is not attractive. We shall consider the problem of determining the sample size necessary to produce a fixed level of precision.

The following eight steps are taken into account when determining the sample size. We will study each one in detail.

(1) Degree of precision desired

(2) Formula to connect n with desired precision

(3) Advance estimates of variability in population

(4) Cost and operational constraints

(5) Expected sample loss due to nonresponse.

(6) Number of different characteristics for which specified precision is

required.

(7) Population subdivisions for which separate estimates of a given precision are required.

(8) Expected gain or loss in efficiency.

5.2 Degree of Precision

The precision of an estimate refers to the amount of variable error, mainly sampling error, contained in an estimate. To lower the sampling error, that is, to increase the precision, we want n to be sufficiently large. Therefore, we decide on a target value for the precision of the estimate. The degree of precision desired can be stated in terms of:

(1) The absolute error (E) for the estimate

$$P[|\hat{\theta}-\theta| \leq E] = 1-\alpha$$

where $\hat{\theta}$ is an estimate of the parameter θ and $(1-\alpha)$ is the degree of confidence desired.

The absolute error E is measured in the same unit used to measure the variable. For example, E = 5 hectares or E = $10,000 or E = 25 persons.

(2) The relative error (RE) for the estimate

$$P[|\frac{\hat{\theta}-\theta}{\theta}| \leq RE] = 1-\alpha$$

This is E expressed as a proportion (or percentage) of the true value of the parameter being estimated. For example, if E = 5 hectares and the true value of the parameter is 100, then RE = 5/100 = 0.05 or 5%.

(3) The target coefficient of variation (cv) for the estimate (v_0)

We set the cv (also known as the relative standard error) for the estimate equal to a target value v_0. For example, we can have:

$$\frac{\sqrt{VAR(\theta)}}{\theta} = 0.05 = 5\%$$

Depending on which of the three ways we use to specify the precision, the formula for n will be different.
The values of E, RE and α are usually decided by the user of the data in conjunction with the statistician.

5.3 Formula that Connects n (sample size) with the Desired Degree of Precision

The following terms are used in the formulas outlined below.

S^2 = the population variance; σ^2 could be used instead.
n = the desired sample size

CV = the population coefficient of variation, ($S/\bar{Y}$) where $\bar{Y}$ is the population mean.

N = Number of units in the population.

k = 1 for 68% confidence
2 for 95% confidence
3 for 99.7% confidence

cv = the coefficient of variation of the estimator ($\bar{y}$, $\hat{Y}$, p, *etc.*)

v_0 = specified target value for estimate's $cv = \dfrac{\sqrt{Var(\theta)}}{\theta}$

E = Absolute error.

RE = Relative error

<u>Note</u>: the level of confidence states the probability that the n determined will provide the degree of precision specified. For example, a 95% level of confidence means that, except for a small chance (5%), we can be 95% certain that the precision specified will be reached with the calculated n. This is equivalent to saying that the acceptable risk is 5% that the true θ will lie outside of the range specified in the confidence interval.

5.3.1. Sample size needed to estimate a mean with absolute error E

The sampling error of a mean using simple random sample is given by (see equation 4.4):

$$S(\bar{y}) = \frac{S}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}}$$

Now, $E = k\ S(\overline{y})$, where k is a multiple of the sampling error, selected to achieve the specified degree of confidence. Therefore, if we substitute $S(\overline{y})$ for (E/k), we get:

$$E = k\ \frac{S}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}} \tag{5.1}$$

If we solve for n in (5.1) above, we get:

$$n = \frac{k^2NS^2}{k^2S^2+E^2N} \tag{5.2}$$

If the population size is large and n ≤ 0.05N, the finite population correction factor in equation (5.1) can be ignored because its effect would be minimal. In this case, we have:

$$n = \frac{k^2S^2}{E^2} \tag{5.3}$$

Example 5.2

Refer to Example 5.1 on page 50. Suppose we would like to estimate P, the proportion of persons of Chinese descent to within ± 3%, with 95% confidence. What sample size do we have to choose to achieve this target? Assume P to be no larger than 1/2.

$$n = \frac{k^2(PQ)}{E^2} = \frac{2^2\ 1/2\ 1/2}{(3\%)^2} = \frac{1}{(0.03)^2} = \frac{1}{0.0009} = 1112$$

Now, let's assume that P ≤ 0.25. What is the required sample size?

$$n = \frac{2^2\ (.25)\ (.75)}{(0.03)^2} = 834$$

Example 5.3

Consider a population consisting of 1,000 farms for which the population variance of the number of cattle per farm is 250 (N = 1,000 and $S^2 = 250$). Let us estimate the average number of cattle per farm from a sample; we wish to have reasonable confidence that the estimate will be close to the true value. Suppose the sample estimate is to be in error by no more than 1 (one head of cattle) from the true average, and we require an assurance of 95 chances out of 100 that the error will be no larger than 1. In this case,

$E = 1 \qquad E^2 = 1 \qquad N = 1{,}000 \qquad S^2 = 250 \qquad N = 1{,}000$

k = 2 (since 2 gives us almost a 95% confidence level); $k^2 = 4$.

Applying equation (5.2), we see that n must be equal to or greater than

$$n = \frac{k^2NS^2}{k^2S^2+E^2N} = \frac{4(1{,}000)(250)}{4(250)+1(1{,}000)} = \frac{1{,}000{,}000}{2{,}000} = 500$$

If in the same situation we are satisfied with an error of not more than 3, with a confidence level of 95 percent, the only change in the formula would be in the values of E and E^2, as follows:

$E = 3$ and $E^2 = 9$.

Then we would have,

$$n = \frac{k^2NS^2}{k^2S^2+E^2N} = \frac{4(1{,}000)(250)}{4(250)+9(1{,}000)} = \frac{1{,}000{,}000}{10{,}000} = 100$$

and a sample of 100 cases would be sufficient.

Example 5.4

We wish to estimate the average age of 2,000 seniors on a particular college campus. How large a SRS must be taken if we wish to estimate the age within 2 years from the true average, with 95% confidence? Assume $S^2 = 30$.

$E = 2$ and $k = 2$

$$n = \frac{k^2S^2}{E^2} = \frac{2^2(30)}{2^2} = 30 \textit{ seniors}$$

5.3.2. Sample size needed to estimate a proportion with absolute error E

The sample size n to estimate a population proportion P is obtained from equation (5.2); in this equation, $S^2 = PQ$:

(5.4) $$n = \frac{k^2NPQ}{k^2PQ+E^2N}$$

And for a large population size ($n \leq 0.05N$), we have from equation (5.3),

(5.5) $$n = \frac{k^2PQ}{E^2}$$

5.3.3. Sample size needed to estimate a total with absolute error E

Using equation (4.3), and letting $E = kS(\overline{y})$, we get the following formula for n:

(5.6) $$n = \frac{k^2N^3S^2}{k^2N^2S^2+E^2N} = \frac{k^2N^2S^2}{k^2NS^2+E^2}$$

If we ignore the fpc, we have:

(5.7) $$n = \frac{k^2N^2S^2}{E^2}$$

5.3.4. Sample size needed to estimate the number of units that possess a certain attribute with absolute error E

To obtain the n necessary to estimate A, the number of units that possess a certain characteristic, simply substitute PQ in place of S^2 in equations (5.6) and (5.7).

5.3.5. Sample size formulas when the error is expressed in relative terms (RE)

We can obtain formulas for estimates when the desired error is expressed in relative terms instead of absolute terms. For relative errors (RE), if (RE) is a proportion of the estimates, substitute (RE/k) for cv($\hat{Y}$) (or $cv(\overline{y})$) in equation (4.7) or (4.8). In order to avoid confusing the estimated coefficient of variation (cv) from the population coefficient variation (CV) we shall denote here the estimated coefficient of variation by cv. We have:

(5.8) $$n = \frac{k^2N(CV)^2}{k^2(CV)^2+N(RE)^2}$$

Note: (RE)/k = cv($\bar{y}$) = s$\bar{y}$($\bar{y}$)/

This applies to both means and totals.

If we ignore the fpc, then equation (5.8) becomes

(5.9) $$n = \frac{k^2(CV)^2}{(RE)^2}$$

NOTE 1: In actual practice, we usually do not know S^2 or $(CV)^2$. Indeed we do not even know s^2 in advance of the survey. Instead, we use rough estimates of S^2 or $(CV)^2$, obtained by the methods discussed in section 8 of chapter 6.

NOTE 2: For the mean and the total, it is better to express the variance in relative rather than absolute terms, for two reasons:

(1) Most importantly, because a population's relative variance is more stable than its absolute variance. A guess or estimate of the population coefficient of variation CV (from past data or from similar populations) is likely to be closer to the true value than a guess or estimate of the variance.

(2) The formula for n is the same for estimators of means or totals when it is expressed in terms of the coefficient of variation.

NOTE 3: To estimate the proportion P, it is preferable to use the absolute error previously discussed because the proportion is itself a relative quantity, so that taking the percentage of a percentage can become confusing.

To obtain the formula for the sample size required to estimate a population proportion when the error is expressed as relative error (RE), use equation 5.8

$$n = \frac{k^2N(CV)^2}{k^2(CV)^2+N(RE)^2}$$

where we replace $(CV)^2 = Q/P$. That is, we get

(5.10) $$n = \frac{k^2NQ}{k^2Q+NP(RE)^2}$$

If we ignore the fpc, equation (5.10) becomes:

(5.11) $$n = \frac{k^2Q}{(RE)^2P}$$

Example 5.5

We would like to carry out a survey to estimate the total area in hectares of the farms in a population. The estimate should be within 10% of the true value. How many farms should be surveyed? (In a pilot survey, we estimated the population coefficient of variation, CV, of the variable farm size to be 1.2). Use 95% confidence.

$$n = \frac{k^2(CV)^2}{(RE)^2} = \frac{2^2(1.2)^2}{(0.10)^2} = 576 \textit{ farms}$$

5.3.6. Sample size formulas when the error is expressed in terms of the coefficient of variation

Equation (5.8) can be expressed in terms of the coefficient of variation. If CV = $\frac{S}{\bar{Y}}$ is the population coefficient of variation and $cv = \frac{(RE)}{k}$ is a specified target value for an estimate's coefficient of variation, then (5.8) becomes,

(5.12) $$n = \frac{\left(\frac{k}{RE}\right)^2(CV)^2}{1+\frac{1}{N}\left(\frac{k}{RE}\right)^2(CV)^2} = \frac{\left(\frac{(CV)}{(cv)}\right)^2}{1+\frac{1}{N}\left(\frac{(CV)}{(cv)}\right)^2}$$

If we ignore the fpc, equation (5.9) gives :

(5.13) $$n = \left(\frac{k}{RE} \right)^2 (CV)^2 = \left(\frac{(CV)}{(cv)} \right)^2$$

Equations (5.12) and (5.13) apply to both means and totals.

Let’s consider Example 5.5 and use coefficients of variation to solve the problem.

Example 5.6

Suppose that a survey was carried out to estimate the total area in hectares of the farms in a population. The estimate should be within 10 percent of the true value, with 95 percent confidence. How many farms should be surveyed? [In a pilot survey, we estimated the population coefficient of variation CV of the variable "farm size" to be 1.2].
In this case,

k = 2 CV = 1.2 (RE) = .10

$$cv = \frac{RE}{k} = \frac{.10}{2} = .05$$

Substituting in equation (5.13), we have,

$$\text{n} = \left(\frac{1.2}{.05} \right)^2 = 576 \quad \textit{farms}$$

Example 5.7

The results from a pilot test are used to estimate $\bar{Y}$ and S for the variable ‘income’ in a population of 5,000 households.

$\bar{y}$ = \$14,852 per household
s = \$12,300

A full scale survey is planned. What should be the sample size for this survey if we want to estimate the mean income per household with a cv no larger than 5%?

The population coefficient of variation CV is estimated by:

$$cv = \frac{12{,}300}{14{,}852} = 0.828 = 82.8\%$$

$$n = \left[\frac{(cv)}{v_0}\right]^2 = \left[\frac{0.828}{0.05}\right]^2 = 275 \;\; \textit{households}$$

5.4. Advance Estimates of Population Variances

In the preceding section, we noted that most of the sample size formulas are written in terms of the population variance. In practice this is unknown and it must be estimated or guessed. There are five ways of estimating population variances for sample size determination.

Method 1: Select the sample in two steps, the first being a simple random sample of size n_1 (the first sample) from which estimates s_1^2 and p_1 of S^2 and P, respectively, are obtained. Then use this information to determine the required n (the final sample size).

Method 2: Use the results of a pilot survey. This is one of the more commonly used methods.

Method 3: Use the results of previous samples of the same or similar population.

Method 4: Guess about the structure of the population and use some mathematical results.

Method 5: (Only for qualitative characteristics.) If the statistic to be measured is a proportion, then make a fairly good guess of P (the proportion in the population).

Method 1 carries out the survey in two steps. In the first step, only a subsample (a random part of the total sample) is enumerated. An analysis of this part permits one to estimate the variance and to make revisions in the total size of the sample, if necessary. In the second step, the remainder of the sample is enumerated in accordance with these changes, if any. This method gives the most reliable estimates of S^2 or P, but it is not often used, since it slows up the completion of the survey.

Method 2 is one of the more commonly used methods. It serves many purposes, especially if the feasibility of the main survey is in doubt. If the pilot survey is itself a simple random sample, the preceding methods apply. But often the pilot work is restricted to a part of the population that is convenient to handle or that will reveal the magnitude of certain problems.

Method 3 is also a very commonly used method. This method points to the value of making available, or at least keeping accessible, any data on standard errors obtained in previous surveys. Unfortunately, the cost of computing standard errors in complex surveys is high, and frequently only those standard errors needed to give a rough idea of the precision of the principal estimates are computed and recorded. If suitable past data are found, the value of S^2 may require adjustment for time changes. Experience indicates that the variance of an item tends to change much more slowly over time than the mean value of the item itself. Even if the mean value changes, the relative error may be quite stable.

Method 4 uses some mathematical results. Deming (1960) shows that some simple mathematical distributions may be used to estimate S^2 from a knowledge of the range (h) and a general idea of the shape of the distribution of the characteristic of interest. $S^2 = 0.0289h^2$ for a normal distribution, $S^2 = 0.083h^2$ for a rectangular distribution (uniform), $S^2 = 0.056h^2$ for a distribution shaped like a right triangle, and $S^2 = 0.042h^2$ for an isosceles triangle.

General Shape of Distribution	Approximate Values of S
Normal	.17h
Equilateral Triangle	.20h
Right Triangle (Skewed Distribution)	.24h
Uniform Distribution	.29h

These relations do not help much if h is large or poorly known. However, if h is large, good sampling practice is to stratify the population (see Chapter 7) so that within any stratum the range is significantly reduced. Usually the shape also becomes simpler (closer to rectangular) within a stratum. Consequently, these relations are effective in predicting S^2, hence h, within individual strata.

Example 5.8

The universities in the State of Maryland were classified according to the number of enrolled students into four size classes. The standard deviation within each class is shown below:

Size Class (i)				
Enrollment Level, X_i	< 1,000	1,000-3,000	3,000-10,000	> 10,000
S_i	236	625	2,008	10,023

If you knew the class boundaries but not the values of S_i, how well could you guess the values by using the Deming method? (No university has fewer than 200 enrolled students and the largest has about 50,000).

We do not know the number of universities in each size class; therefore, we cannot obtain a frequency distribution that would show us the general shape of the distribution. A conservative estimate would be that the distribution is uniform. In this case, S_i would be given by $0.29*h_i$, where h_i is the range of each class.

$$
\begin{array}{llll}
S_1 & = & 0.29\,(1{,}000 - 200) & = & 232 \\
S_2 & = & 0.29\,(3{,}000 - 1{,}000) & = & 580 \\
S_3 & = & 0.29\,(10{,}000 - 3{,}000) & = & 2{,}030 \\
S_4 & = & 0.29\,(50{,}000 - 10{,}000) & = & 11{,}600
\end{array}
$$

Method 5: if the statistic to be measured is a proportion--for example, the proportion of farms growing corn--the population variance is PQ. It is only necessary to be able to make a fairly good guess at P in order to estimate S^2. As long as the guess is reasonably close, we will get a good estimate of S^2. For example, suppose the true value of P is 0.4; then the value of $S^2 = PQ$ would be 0.4 x 0.6 = 0.24. Suppose we made a rather poor guess and P, such as 0.3. We would then estimate the value of the variance as 0.3 x 0.7 = 0.21, which differs from the true value by only about 10 percent. Note that we can also estimate S^2 by setting $S^2 = PQ = (1/2)(1/2)$ because the formula for n is maximized when $P = Q = 1/2$. This latter is called a "conservative estimate," because we can never do worse than that.

5.5. Cost and Operational Constraints

Let us recall that the total error (inverse of accuracy) is composed of both bias and variance. High sample sizes reduce the variance (i.e., yield high precision) but tend to increase cost and operational difficulties, which translates into larger nonsampling errors.

To reduce the incidence of nonsampling errors, a survey needs:

(1) good quality control
(2) sufficient resources.

However, in a real survey setting, there exist constraints with respect to:

(a) budget
(b) field conditions
(c) field and office personnel
(d) time
(e) equipment and materials, etc.

Hence, in addition to precision, we also need to consider the maximum sample size that can be handled by the available resources. It may be necessary to limit the sample size in order to stay within budget and operational constraints.

If the maximum practical sample size is much smaller than that required to achieve the specified precision, calculations can be made to estimate the level of precision that could be expected from the actual sample size. If this level is not acceptable, greater resources have to be allocated to accommodate a larger sample size.

To compromise between precision and practicality, we may take a sample size that is somewhere between the constraint-based and the precision-based sizes.

5.6. Expected Sample Loss Due to Nonresponse

If past experience indicates that a certain level of nonresponse can be present, we may want to inflate the calculated sample size to compensate. This is because our calculations were based on a 100 percent response. If we do not obtain all the interviews, then the estimates will be based on a number smaller than the calculated n and will, therefore, have a greater variance than expected.

Inflating Procedure

We compute the inflated sample size n' from the following relationship:

$$n' = \frac{n}{r}$$

where r is an estimate of the expected response rate and it can be obtained from previous rounds of the same survey, previous experience with similar surveys, a pilot (pre-test), etc.

For example, we calculate n to be 1,000 units. Based on the results of a pilot survey, we anticipate the response rate to be 70 percent.

Our inflated n will be: n' = 1000/.70 = 1,429

If our assumption was correct, we should get back 70% of 1,429 = 1,000.

Therefore, our estimates will be based on the same number of units as expected and the target precision will be attained.

Important Note

Inflating the sample size when there is nonresponse only helps compensate for the resulting loss in precision. It does nothing for diminishing the resulting nonresponse bias.

5.7. Number of Different Characteristics Requiring a Specified Precision

In most surveys information is collected from a sampling unit for more than one characteristic. One method of determining sample size is to specify margins of error for the characteristics that are regarded as most vital to the survey. An estimation of the sample size needed is first made separately for each of these important characteristics.

When the estimations of n have been completed for each of the most important characteristics, it is time to take stock of the situation. It may happen that the n's required are all reasonably close. If the largest of the n's falls within the limits of the budget, this sample size is selected. More commonly, there is sufficient variation among the n's so that we are reluctant to choose the largest, either for budgetary considerations or because this will give an overall level of precision substantially higher than originally contemplated for the other characteristics. In this event the desired level of precision may be relaxed for some of the characteristics in order to permit the use of a smaller value of n.

In some cases the n's required for different characteristics are so different that some of them must be dropped from the survey; with the resources available the precision expected for these characteristics is totally inadequate. The difficulty may not be merely one of sample size. Some characteristics call for a different type of sampling scheme than others. With populations that are sampled repeatedly, it is useful to gather information about those characteristics that can be combined economically in a general survey and those that need special methods. As an example, a classification of characteristics into four types, suggested by experience in regional agricultural surveys, is shown in Table 5.1. In this classification, a general survey means one in which the units are fairly evenly distributed over some region as, for example, by a simple random sample.

Table 5.1.

AN EXAMPLE OF DIFFERENT TYPES OF ITEMS IN REGIONAL SURVEYS

Type	Characteristics of item	Type of Sampling Needed
1	Widespread throughout the region, occurring with reasonable frequency in all parts.	A general survey with low sampling fraction.
2	Widespread throughout the region but with low frequency.	A general survey, but with a higher sampling fraction.
3	Occurring with reasonable frequency in most parts of the region, but with more sporadic distribution, being absent in some parts and highly concentrated in others.	For best results, a stratified sample with different intensities in different parts of the region (Chapter 5). Can sometimes be included in a general survey with supplementary sampling.
4	Distribution very sporadic or concentrated in a small part of the region.	Not suitable for a general survey. Requires a sample geared to its distribution.

Example

The following coefficients of variation per unit were obtained in a farm survey in Iowa, the unit being an area 1 square mile.

Item	Estimated cv
Acres in farms (Y_1)	0.38
Acres in corn (Y_2)	0.39
Acres in oats (Y_3)	0.44
Number of family workers (Y_4)	1.00
Number of hired workers (Y_5)	1.10
Number of unemployed (Y_6)	3.17

A survey is planned to estimate acreage characteristics with a cv of 2½% and numbers of workers (excluding unemployed) with a cv of 5%. With simple random sampling, how many units are needed? How well would this sample be expected to estimate the number of unemployed? The results are displayed in the following table:

Item	Estimated cv	Target cv for Estimate	n	Expected cv[9] with n = 484
Y_1	0.38	0.025	232	0.017
Y_2	0.39	0.025	244	0.018
Y_3	0.44	0.250	310	0.020
Y_4	1.00	0.050	400	0.046
Y_5	1.10	0.050	484	0.050
Y_6	3.17	--	--	0.144

Comments

1. Assuming cost and workload constraints permitted it, a sample of 484 segments should be taken (the largest calculated size). This sample size should guarantee the desired precision (or better) for the estimates of Y_1 through Y_5. As noted in the last column, the cv of the estimate is expected to be either as small as desired or smaller, if n = 484 is used.

2. As far as the estimate of Y_6, a cv of approximately 14% can be expected if a sample size of 484 is used. Although it is true that the precision will be lower for this estimate than for the others, this is not critical because sponsors and data users did not require higher precision.

5.8. Population Subdivisions Requiring Separate Estimates of a Specified Precision

If there are subpopulations or domains of estimation for which separate estimates of a given precision are required, we must resort to a different sampling strategy, such as the use of stratified sampling with different sampling rates by stratum.

Under stratified sampling, each stratum or domain is considered a "population" in its own right. We can then apply the same principles to calculate separate sample sizes within each stratum to

[9] $CV(\hat{Y}) = \dfrac{s(\hat{Y})}{N\bar{y}} = \dfrac{\frac{Ns}{\sqrt{n}}}{N\bar{y}} = \dfrac{\frac{s}{\bar{y}}}{\sqrt{n}} = \dfrac{cv}{\sqrt{n}}$

meet the precision requirements for the domain estimates. Often the same precision is required in each domain. If the variability and the cost within the domain are similar from domain to domain, then the sample sizes will be about the same in all domains.

The overall sample size would then be the sum of the stratum sample sizes. The overall estimate for the whole population would have a higher precision than the stratum-level estimates.

For example, if the unemployment rate is to be measured at the national level with x% target cv, the national sample size computed would be n, say 5,000 households. On the other hand, if the unemployment rate is needed for each of 5 regions of the country, all with the same precision, the total (national) sample size required would be around 5n or 25,000 households. The national estimate would have a precision much higher than originally planned.

5.9. Expected Gain or Loss in Efficiency

The formulas discussed so far are all based on simple random sampling (SRS). Let us denote as n_{srs}, the sample sizes obtained from those formulas.

However, as will be seen later on, simple random sampling is rarely used in complex surveys. The efficiency of the design actually used is measured by comparing the variance of the estimator Θ obtained with the complex design and the variance of the same estimator with SRS.

- If the complex design is more efficient, that is, inherently tends to produce a lower variance than SRS, then our precision is likely to be better than expected with n_{srs}.

- If, on the other hand, the complex design is less efficient than the SRS one, that is, has an inherent tendency to produce a higher variance for Θ than SRS, then our expected precision level may not be met with the calculated n_{srs}. In this case, it would be desirable to inflate n_{srs} beforehand.

As we study different sampling schemes, we will know which are more efficient than SRS and which are less. Here are some examples:

Usually more efficient than SRS:

- stratified sampling, implicit stratification in systematic selection, use of more efficient estimators (e.g., ratio estimators of total)

Generally less efficient than SRS:

- cluster sampling (used for convenience and cost effectiveness)

The efficiency of a particular sample design is measured by the design effect (see Chapter 4).

5.10. Relationship Between Size of Sample and Size of Population

We return to certain implications of the basic formula from which all the above formulas are derived. That basic formula was given in equation (4.4) in Chapter 4 as:

(5.14) $$S^2(\bar{y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right)$$

Notice that the sampling variance of the mean is equal to the variance of individual observations (S^2) in the population multiplied by the factor $\frac{1}{n}\left(\frac{N-n}{N}\right)$. What happens when the sample increases from its smallest possible size (n = 1) to its largest possible size (n = N)? When n = 1,

$$S^2(\bar{y}) = \left(\frac{N-1}{N}\right)\frac{S^2}{1} \doteq S^2$$

This states the familiar fact that the variance of the means of samples of one unit is the same as the variance of individual observations in the population. At the other extreme, when n = N,

$$S^2(\bar{y}) = \left(\frac{N-N}{N}\right)\frac{S^2}{N} = 0$$

That is, if the sample includes the entire population, the mean is estimated without sampling error.

For sample sizes between these extremes, how does the sampling fraction (sampling rate) n/N affect the standard error? The answer, sometimes surprising to students, is that for populations that are large relative to the sample size, the <u>absolute size of the sample</u> (n) and not the sampling fraction n/N <u>determines the precision of the estimated mean</u>. This follows from the fact that when N is large relative to n, the factor [(N-n)/N] ≈1. (The symbol ≈ stands for "is approximately equal to."). Then $S^2(\bar{y}) \doteq \frac{S^2}{n}$; thus, it is clear that the error depends on S^2 anc

n, and not on $\frac{n}{N}$.

On the other hand, for small populations the sampling fraction <u>does</u> have an effect. For example, suppose two populations have the same mean and the same variance: $\bar{Y}$ = 50 and S^2 =

100, while $N_1 = 40$ and $N_2 = 400$. If we take the same size of sample from each, say $n_1 = n_2 = 20$, the standard errors <u>are</u> related (in an inverse way) to the sampling fractions. Equation (5.14) then gives:

	$\underline{N}$	$\underline{n}$	$\frac{n}{N}$	$S(\bar{y})$
1st population	40	20	.50	1.6
2nd population	400	20	.05	2.2

The number of sample units needed to achieve the same precision would be greater for the second (larger) population. However, the number of sample units needed to achieve a given reliability does not increase indefinitely as the number of elements in the population increases. In other words, we reach a point in which adding an extra sampling unit does not produce a sizable reduction in variance.

<u>Example</u>

Table 5.2 below shows the size of sample necessary to give an estimate of the population mean within a 5 percent error (E = 0.05) of the estimate (with confidence coefficient k = 2) for populations ranging in size from 50 to 10,000,000 elements and with $CV^2 = .10$ in each case. These results were obtained using equation (4.8) of Chapter 4. Equation 4.8 is given by:

$$CV(\bar{y}) = \frac{S(\bar{y})}{\bar{y}} = \frac{\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N}}}{\bar{y}} = \frac{CV}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$$

Table 5.2

NUMBER OF ELEMENTS NECESSARY FOR FIXED PRECISION: $CV^2 = .10$
(E = .05 and k = 2)

–

Number of elements in the population (N)	Number of elements required in sample (n)	n/N

50*.....................	38	.76
100.....................	62	.62
1,000.................	138	.14
10,000................	158	.016
100,000.............	160	.0016
1,000,000...........	160	.00016
10,000,000.........	160	.000016

* Use equation (4.8) when N is smaller than 50.

As an example, let's calculate the first value of n in Table 5.2. Since N = 50 and is very small for a population value, we have to use the formula for n that contains the finite population correction factor (N-n)/N. The series of steps leading to the number 38 in Table 5.2 is shown below.

$$CV(\bar{y}) = \frac{S(\bar{y})}{\bar{y}} = \frac{CV}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$$

$$[CV(\bar{y})]^2 = \frac{(CV)^2}{n}\frac{N-n}{N}$$

$$[CV(\bar{y})]^2 = \frac{(CV)^2}{n} - \frac{(CV)^2}{N}$$

The objective is to leave n on one side of the equation in terms of the other components.

$$[CV(\bar{y})]^2 + \frac{(CV)^2}{N} = \frac{(CV)^2}{n}$$

Now, we know that $(CV)^2 = .10.$ This is the population coefficient of variation and is given to us as a known value. However, we do not know the value of $[CV(\bar{y})]^2$ but we can obtain it by using the following:

$$n = \frac{(CV)^2}{[CV(\bar{y})]^2 + \frac{(CV)^2}{N}}$$

$$CV(\bar{y}) = \frac{s(\bar{y})}{\bar{y}} = \frac{E/k}{\bar{y}} = \frac{5\%\bar{y}\ /2}{\bar{y}} = \frac{0.05}{2} = 0.025$$

Consequently, we have the following value for n when N = 50:

$$n = \frac{.10}{0.000625 + \frac{.10}{50}} = \frac{.10\ (50)}{50\ (0.000625) + .10} = \frac{5}{.13125} \approx 38$$

Table 5.2 shows that for small populations, the sample size needed for a given accuracy does increase as the population increases, but the sample size approaches a fixed number as the population gets very large. The largest size of sample we would ever need for this accuracy (with $CV^2 = .10$) is 160 elements, and this is approximately the number we would need whether there are 10,000 or 10,000,000 elements in the population. Furthermore, if we had used a sample of 160 for a population even as small as 1,000, the sample would be somewhat larger than necessary; but the excess would not have been very serious.

Study Assignment

Problem A. *You want to design a household survey to estimate average annual income per household. The number of households is 2,000,000. On the basis of the data from a previous census, the population variance of annual income per household is estimated to be 1,000,000 (that is, S = 1000).*

Exercise 1. *What sample size is necessary to estimate the average annual income with a 95 percent confidence that the result is accurate to plus or minus $100?*

Exercise 2. *What size sample is necessary to estimate average annual income within plus or minus $50, also at the 95 percent confidence level?*

Exercise 3. *We wish to estimate the average age of 3,000 seniors on a particular college campus. How large a simple random sample must be taken if we wish to estimate the age within 2 years from the true average, with 95 percent confidence. Assume S^2 = 30 and ignore the fpc.*

Problem B. *A survey is to be made of the prevalence of common diseases in a large population. For any disease that affects at least one percent of the individuals in the population, it is desired to estimate the total number of cases with a coefficient of variation of not more than 20 percent.*

Exercise 4. *What size of simple random sample is needed, assuming that the presence of the disease can be recognized without mistakes?*

Exercise 5. *What size is needed if total cases are wanted separately for males and females with the same precision?*

Chapter 6.

PRACTICAL CONSIDERATIONS IN SELECTING A SAMPLE

6.1 SAMPLING FRAME

In order to select a sample, it is necessary to have a sampling frame; that is, a list of all elements (or the equivalent, such as a list of blocks, housing units, etc.) so that the probability of selection of each element can be known in advance. The frame need not be literally a list. In sampling from cards, questionnaires, etc., the documents themselves can be considered as the frame. But it is necessary to know that the file is complete. For example, in sampling from a file of records, one should make sure that no records are out of the file--in use or waiting to be refiled--since such records would not have any chance of selection. Again, in using a population register maintained by local authorities, one should make certain the list is current. For example, the list might not contain all families with married couples. Since new families and those that move around are likely to differ in their characteristics from older and more settled families, a biased sample would result.

In using local registers or lists, it may be useful to conduct an actual check of the completeness, on a more or less informal basis. This can be done by going out to the area to be sampled, selecting a few families (or farms or business firms) scattered around the area, and checking to see if they are on the list. If possible, it is better to select families of the type likely to be missing from the list, since this would provide a better test. A rough idea of the adequacy of the list can be obtained in this manner.

6.2 PROBABILITY OF SELECTION OF UNITS

Special difficulties arise when some units have more than one chance of selection--for example, when sampling from a file in which some individuals are included more than once; when selecting a sample of families from a sample of individual persons; etc. To illustrate, one might select a sample of school children and use it to select families. It is clear that if one draws a sample of families by first selecting a sample of persons and including the families to which these persons belong, the families will have unequal probabilities of selection, since the larger the family the greater the chance of selection. Similarly, selecting a sample of a business firm's customers by using a record file containing a separate sheet (or card) for each purchase will give customers making more than one purchase a greater chance of selection.

To avoid the biases which result from giving some of the units a greater chance of selection than others, it is desirable to restrict the sampling procedure so that each unit has only one chance of selection. For example, when selecting a sample of families, we could make a rule to include the family only if the head of the family is the person selected. Since each family has only one head, each family would have the same chance of selection.
The specified person on whom the selection of the family depends need not be the head; he/she could just as well be the oldest person, the youngest child, etc. The only requirement is that each family have one and only one such member. Similarly, in sampling customers, we could restrict the sample by using only the cards with the earliest date for each customer, etc.

While the technique described in the preceding paragraph is generally recommended, whether the sample is drawn from a file, a set of questionnaires, or is selected in the field, there are other techniques that might be used. They will provide unbiased estimates of the universe, although they do not strictly satisfy the conditions of simple random sampling. Some of these techniques are:

(1) After selecting the initial sample by including all families for which one (or more) person has been selected, we group the sample by size of family. It is clear that families with 2 members have twice the chance of selection as those with 1; families with 3 members have three times the chance of selection; etc. Therefore, instead of interviewing all families in the sample, we interview only 1/2 of the two-member families; 1/3 of the three-member families; etc.

(2) Proceed as above, but interview all families instead of 1/2, 1/3 etc. However, in tabulating the results, tabulate each size class separately, and multiply the results of the two-person families by 1/2, the three-person families by 1/3, etc., before adding the results together.

6.3 FRAMES INCLUDING OUT-OF-SCOPE UNITS

Sometimes the only available frame is a list which includes some units which are outside the scope of the universe defined for the survey. For example, suppose a special analysis is desired of the census characteristics of males. The only source for sampling is a card file containing cards for all persons both male and female, and it is not feasible to remove all the cards for females. The file can still be used as a frame even though cards for both males and females will be designated by the random selection process. The proper procedure in such a case is to take only the cards for the males selected, and disregard those for the females.

Do not substitute. A procedure that is sometimes erroneously used (and may cause serious bias) is to substitute the next "male" card in the file for each "female" card drawn in the sample. There are two things wrong with this method:

(1) It results in a higher sampling rate than that specified. Also, the sampling rate actually obtained cannot be calculated unless the total number of males is known. This makes it impossible to use the reciprocal of the sampling rate, N/n, as a multiplier to produce estimates of totals from the sample.

(2) A more serious objection to this substitution lies in the biases it may introduce in the selection process. Suppose we have a list of all housing units and we wish to select a sample of occupied dwellings only. If we use a procedure that substitutes the next occupied unit for each vacant housing unit that falls into the sample, occupied units that are neighbors of vacant ones will have two chances of selection--the chance that their own listing entry is selected and the chance that the listing of the neighboring vacant dwelling is selected. If vacant units are more likely to be found in poor and undesirable neighborhoods, this would mean that occupied housing units in such areas would be over-represented in the sample.

6.4 SYSTEMATIC SAMPLING

The work necessary to draw a simple random sample can be quite burdensome when the number of units to be selected is large. For example, to get a 5 percent sample of 20,000 elements, it would be necessary to select 1,000 random numbers from a table of random numbers and then to select the designated units from the population. In practice, most statisticians prefer a different method. A sample of this size is usually drawn by taking a random number between 1 and 20, then taking every 20th element thereafter. Thus, if the random number is 3, the elements taken will be 3, 23, 43, 63, and so on up to 19,983. The reciprocal of the sampling rate (20 in this case) is called the sampling interval. The method of estimating the mean, total, or a proportion is the same as for simple random sampling.

This type of sampling is called systematic sampling. It is not the same as simple random sampling, but it is an acceptable sampling method because the chance of selecting any one element is known and we can calculate the sampling errors.

If the elements in the population are arranged in a nearly random order (that is, with very little correlation between successive elements), the results of systematic sampling will be in close agreement with those of simple random sampling. Experience shows that, generally, the two methods will give results of roughly the same accuracy. The systematic sample will often have a somewhat smaller sampling error, since it will make certain the sample will be spread throughout the population. We may make use of the formulas for simple random sampling to evaluate the reliability of estimates from a systematic sample; the result will usually somewhat overstate the standard error for systematic sampling. In other words, we will underestimate, slightly, the reliability of the estimates. There are ways of calculating the standard errors of systematic samples more precisely; however, they are not covered in these chapters.

6.4.1 General Procedure for Selecting a Sample

The systematic sample selection procedure consists of the following steps:

1. Assign serial numbers from 1 through N to the population units.

2. Calculate SI = N/n, the sampling interval

   - for exactness, carry as many decimals as possible
   - you may round if you are doing this without a calculator, but you would be sacrificing exactness for convenience

3. Select a random number (RN) from a table of random numbers between 0 and the SI. This is called a random start (RS)

   - in the permitted range, exclude zero, but include the sampling interval
   - use as many digits as SI has, including decimals
   - if you are searching through a RN table, pretend the decimal point is not there
   - if you are using a calculator which only provides random numbers between zero and one, multiply this random number by the value of SI in order to get a random number between zero and SI. Remember to keep the

decimals, do not round yet.

4. Begin the series of cumulated numbers with RS. Add SI to this first number to determine the second. Then, add SI to the second number to get the third, and so on.

    - Do not round decimals during the addition process

5. Stop cumulating when the last cumulated number exceeds N (discard this last number)

    - this should occur when you have cumulated n numbers
    - if you rounded SI before adding, you may not have exactly n

6. Now go back and round all the cumulated numbers up to the next integer

7. On the list of population units, circle the serial numbers that correspond to these integers

    - These are the selected units.

Example 6.1

Suppose that a village contains 285 housing units (HUs) and we wish to select a systematic sample of 12 HUs for a survey. Assume the list is randomly ordered.

We want to determine the HUs that will be in the sample.

1. SI = N/n = 285/12 = 23.75

2. RN between 0001 and 2375 is 1979

3. RS = 19.79

4. Series of cumulated numbers:

Sample Unit	Selection Number	Actual Unit Selected

	of Selected Unit	(Serial Number after rounding up)
1	19.79	20
2	19.79+23.75=43.54	44
3	43.54+23.75=67.29	68
4	67.29+23.75=91.04	92
5	114.79	115
6	138.54	139
7	162.29	163
8	186.04	187
9	209.79	210
10	233.54	234
11	257.29	258
12	281.04	282
13	304.79	(Discard)

Remarks: Let's see what might have happened if we had not carried the decimals.

1. SI = N/n = 285/12 = 23.75 rounded up to 24.

2. Suppose RN between 01 and 24 is actually 24.

3. RS = 24

4. Results:

(1) 24
(2) 48
(3) 72
.
.
(11) 264
(12) 288 (discard).

We exhausted the population before reaching our 12 units. This would not have happened if we had kept the decimals (had not rounded up at the beginning), even if our RN was equal to the SI.

6.4.1.2 Useful Variation for Use with Computer Software Packages

We accomplish the same results by truncating instead of rounding up. Refer to Section 4.1 above.

- In step 3 of Section 4.1, while choosing RN, include zero but exclude SI;
- Add 1 to RN to define RS;
- Then, in step 6, truncate (that is, retain only the integer portion of the number), instead of rounding up.

This alternative is convenient when using computer software packages because their rounding functions usually round up to the closest number instead of up systematically. So, it is better to use the integer functions which truncate systematically.

Let's look at an example in order to clarify the concepts. Refer to the previous example.

1. SI = N/n = 285/12 = 23.75

2. RN between 0000 and 2374 is 1979.

3. RS = 19.79 + 1 = 20.79

4. Series of cumulated numbers:

Sample Unit	RN + k (SI)	Actual Unit Selected (Serial Number after truncating)
1	20.79	20
2	20.79+23.75=44.54	44
3	44.54+23.75=68.29	68
4	68.29+23.75=92.04	92
5	115.79	115
6	139.54	139
7	163.29	163
8	187.04	187
9	210.79	210
10	234.54	234
11	258.29	258
12	282.04	282

13	305.79	(Discard)

6.4.2 Caution in the use of systematic sampling

There is one situation in which systematic sampling will give very poor reliability. That is the case in which the arrangement of the elements in the population follow a very regular (periodic) pattern and the sampling interval of the systematic sample falls into that pattern. For example, suppose all families in a certain population consisted of exactly four persons--the head, his wife, and two children. The population has been listed in the order just given and we wish to draw a 25 percent systematic sample from this list to obtain some special information. Since the sampling procedure is to take every fourth person starting at random, four possible samples could be obtained:

(1) Random start is 1--the sample will consist entirely of heads of families.

(2) Random start is 2--the sample will consist entirely of wives of heads.

(3) Random start is 3 or 4--the sample will consist entirely of children.

In a case such as this, results from sample to sample would have nearly the maximum possible variation, and it would be likely that estimates based on any one of the samples would be quite far from the true values for the population. However, even in this extreme case, the estimates would be unbiased; that is, the averages of the estimates for all possible samples would be the population averages.

Although the example given above is not likely to occur in practice, approximations to this situation sometimes arise. If there is suspicion of any regularity in the sequence of listing, which could conform to the sampling interval, systematic sampling should be avoided or modified. For example, the list could be randomized before systematic selection is used.

6.4.3 Modified systematic sampling

One variant of systematic sampling that could be used when there is some systematic ordering in the population is to use a different random number within each sampling interval. To illustrate, let us use the previous example of 25-percent sample when family members are listed in order--head, wife, child. With a systematic sample, once a random number is selected, this sets the pattern for the entire sample. As explained above, if the random number is 1, the sample will be the 1st, 5th, 9th, 13th person, etc. (all heads of families); if the random number is 2, the sample will include the 2nd, 6th, 10th, 14th person, etc. (all wives of heads). To avoid this difficulty, we can select a different random number within each group of 4 persons, so as to avoid a constant interval between our sample cases. The selection scheme is indicated below:

Random number (1 to 4)	Group of four persons	Person selected
3	1st	3rd
1	2nd	5th
2	3rd	10th
1	4th	13th
4	5th	20th
etc.		

That is, in the first group, one child is selected because the random number is 3. In the second group, the husband is selected because the random number is 1 and the husband is the first person in the group, but the fifth person in the list. In the third group, the second person is selected (the wife), who is the 10th person in the list, and so forth.

The system requires more work than ordinary systematic sampling, but it avoids the possibility of the patterns indicated above. We do not mean to imply that such patterns as described above usually exist and that systematic sampling should be avoided. In most cases, systematic sampling produces very satisfactory results.

6.4.4 Serial number as a sampling source

Frequently, in sampling office files, the records have a serial number. We may take advantage of this fact to draw the sample; for example, by designating all records whose serial numbers end in 5, 7, or some other number chosen from a table of random numbers. However, before deciding on this system, one should make sure that the last digit of the serial number is actually random, and does not represent a nonrandom arrangement of some kind; if it does, we might obtain only one particular type of unit in the sample by repeatedly selecting the same last digit. If such a serial number is not present, frequently one can be assigned at random with little cost, and used for sampling.

6.5 GUIDELINES ON WHEN TO USE DIFFERENT SAMPLING SCHEMES

6.5.1 When to Use Simple Random Sampling (SRS)

Some situations which suggest the use of SRS are:

1. There are no major cost differences associated with including various classes of sampling units in the sample.

2. The population is relatively homogeneous with respect to the major characteristics being estimated.

3. There is no auxiliary information available for the population units.

4. There are no cost savings in surveying units which are close together or other natural clusters of the population.

5. A sampling frame which lists each population element is available.

6. There is no need to make separate estimates for subdivisions of the population.

It should be noted that none of these reasons on its own is enough to justify the use of SRS.

6.5.2 When to Use Systematic Sampling

There are several reasons for using systematic sampling, but in practice, the main reason usually is:

- to select a SRS quickly (from a randomly ordered list)

This type of systematic sampling is suggested for SRS when:

1. The frame is a record system requiring a manual selection of sample units (e.g., a physical list, card files, etc.)

2. Sampling units are arranged in random order.

3. Time and resources for selecting the sample are limited.

4. No periodicity is suspected in the data.

Systematic sampling can also be used to provide implicit stratification during sample selection if sampling units are arranged in a particular order. This type of sampling, however, would not be SRS.

6.5.3 When to Use Stratification

Some situations which suggest the use of stratified sampling are:

1. Natural or predefined strata of the population exist: e.g., geographic divisions such as states, provinces; ecological zones that have great socioeconomic impact on the population, etc..

2. There exist subpopulations of interest for which separate estimates of a given precision are required.

3. For administrative convenience, such as regional offices of national statistical offices. Strata could be created so that each regional office can handle the sampling and the interviewing in their respective areas.

4. Stratification can provide a reduction in cost.

5. Stratification can provide a reduction in variance. This would occur if

    a. The variables of interest are correlated with the variable of stratification.

    b. The potential strata are internally homogeneous with respect to the variables of interest.

6. Auxiliary information upon which to base the stratification is available for all population units.

7. Different sampling strategies are required in different parts of the population.

6.5.4 When to Use Single-Stage Cluster Sampling

Some situations which suggest the use of cluster sampling are:

1. Natural or predefined clusters of the population exist: e.g., Metropolitan Statistical Areas (MSAs), Enumeration Districts (EDs), Enumeration Areas (EAs), etc.

2. Confining sampling operations to units that are nearby produces large cost and time savings.

3. No frame is available which lists all population elements but one could be constructed for a limited number of clusters to list all elements in the cluster.

4. Elements within clusters are heterogeneous with respect to variables of interest.

5. Cluster means of the variables of interest are similar among themselves.

6. Cost savings justify the relative loss in precision.

7. Nonsampling errors can be controlled more effectively (e.g. listing operation can be done more accurately for a cluster than for the whole population, yielding better coverage).

It is generally recommended that clusters be selected either with probability proportional to size or with equal probabilities after stratification by size. In addition, it is recommended that larger clusters be placed in certainty strata so they may all be included in the sample. This is done in order to control the variance of estimates.

6.5.5 When to Use Multi-Stage Sampling

The situations which suggest the use of a multistage design are the same as for single stage cluster sampling except that multistage sampling is preferred over single stage sampling when:

1. It is operationally impractical to survey all elements in a cluster, or

2. Only a limited number of sample elements can be handled, and concentrating them in a few clusters would result in estimates of poor precision. In such a case, it would be more efficient to spread the sample over more clusters and only subsample each cluster.

Remarks: The above guidelines for using different sampling schemes are not meant to be rigid or exhaustive. In practice, there might be:

- multiple survey objectives that conflict with one another, or

- survey objectives which conflict with survey resources.

Hence, it is usually necessary to compromise in selecting a design or often to combine designs.

6.6 CONTROLS

After a sample is selected, it is necessary to check the number of cases actually obtained against the number expected (as calculated by applying the sampling rate to the number of cases in the universe). Discrepancies may indicate that the sampling procedure was not properly carried out. For example, forgetting to sample from file drawers in use at the time of sampling, and thus omitting part of the population, would result in fewer cases than expected. Further checks on whether the sample shows any unusual features may also help us know whether the sampling was actually performed as planned.

6.7 USE OF CHECK DATA IN SAMPLING

Very frequently, when a sample has been selected for a study, sample data will be collected and tabulated for a set of basic items for which there are already available known population totals in addition to the items of special interest in the survey. Such known population totals are called "check data" or "independent information." If the sample results for the known items agree closely with the known population totals, it is sometimes claimed that this coincidence "validates" the sample and proves it will provide good results for other items.

Actually, this so-called "validation" does not demonstrate that we have a "good" sampling procedure, or that the sample will yield "good" estimates for the other items in the survey. It is only on the basis of a random method of selecting the sample that we are able to attach a sampling error to our statistics, and to evaluate the probability that the estimates will be within specified limits of the true value: therefore, it is obvious that we cannot rely exclusively on such "validation."

Nevertheless, there are three acceptable uses of check data:

(1) Available check data may be used in improving the method of sampling; for example, in providing a basis for stratification. (This is the subject of the next two chapters.)

(2) It is possible to calculate the standard errors of the estimates made from the sample data. If the check data and sample estimates of the same items differ more than might reasonably be expected from the size of the calculated standard errors, this may indicate that the sampling procedures may not have been carried out properly, the sampling frame has coverage errors, or something else may have gone wrong in the implementation of the survey. Further investigation is needed.

(3) Check data may be used in improving the method of making estimates from the sample; for example, by adjusting the sample estimate by the ratio of the true value of the check item to the sample estimate of this check item (using a ratio estimate). We will discuss this more fully in later chapters.

The above three applications of the use of check data (or independent information) are acceptable, since we can make statistical inferences when using them.

Study Assignment

Problem A: *You have a population of 185 persons. You want to select a sample from this population*

Exercise 1. *Select a systematic sample of 20 persons. List the numbers assigned to them and describe the procedure you used in the selection.*

Problem B: *Suppose there are 25,000 housing units (HUs) in a city, and an accurate listing of the HUs is available. The housing units are listed alphabetically by the family name of the occupants and the addresses also are given. You wish to make a sample survey of rented housing units to estimate the distribution of monthly rentals. You have decided that a total sample of 400 rented housing units is needed to give fairly accurate data. The list of housing units in the city does not show whether the housing units are occupied by renters or owners, but you know that about two-thirds of all the housing units in the city are rented.*

Exercise 2. *Describe how you would draw the sample, including (a) the method of selection, (b) the sampling rate, and (c) the treatment of housing units occupied by owners.*

Exercise 3. ***Suppose that a city block contains 125 housing units. We wish to select a systematic sample of 10 housing units. Follow the steps we discussed in section 4.1 to accomplish this.***

1. SI = -------

2. RN between ------ and -------- is

3. RS = --------

4. Series of cumulated numbers:

Sample Unit	***Selection number of of selected unit***	***Actual unit selected (Serial number after rounding up)***
1	------	------
2	------ + ------ = ------	------
3	------ + ------ = ------	------
4	------ + ------ = ------	------
5	------ + ------ = ------	------
6	------ + ------ = ------	------
7	------ + ------ = ------	------
8	------ + ------ = ------	------
9	------ + ------ = ------	------
10	------ + ------ = ------	------

Exercise 4. *Select a systematic sample of 100 housing units from a population of 1250 housing units. Set up the table showing the sampling unit, selection number and actual unit selected.*

Chapter 7

STRATIFIED SAMPLING-BASIC THEORY

7.1 DESCRIPTION OF THE STRATIFICATION PROCEDURE

In simple random sampling, we do not try to force the sample to be representative of different groups in the population. The tendency to be representative is inherent in the procedure itself and the sampling error can be reduced only by increasing the size of sample. However, if something is known in advance about a population, it may be possible to use this information in stratification and thus reduce the sampling error. The judgment of experts may be useful here.

Stratified random sampling is a method in which the elements of the population are divided into groups (strata), and a simple random sample is selected for each group, taking at least one element from each group (stratum). One element from each group is sufficient to estimate the mean, but two are needed to estimate its reliability; generally many more than two are needed to make the estimates sufficiently precise. The process of establishing these groups is called stratification and the groups are called strata. The strata may reflect regions of a country, densely populated or sparsely populated areas, various ethnic or other groups.

In stratification we group together elements which are similar, so that the variance S_h^2 within stratum h is small; at the same time, it is desirable that the means of the several strata $(\bar{Y}_h)$ be as different as possible. The letter h will be used to identify the strata so that if L strata are created, h will go from 1 to L.

In stratified sampling, the probabilities of selection may be the same from group to group, or they may be different. It is not necessary that all elements have the same chance of selection, but the chance of each must be known. Under stratified random sampling all the elements in a particular stratum have equal chances of being selected.[10] While not every combination of elements is possible, all of the possible samples (that is, combinations of elements) that might be drawn have the same chance of occurring.

In stratified sampling, the selection of sampling units, the location and enumeration of the selected units, distribution and supervision of fieldwork and, in general, the whole administration of the survey is greatly simplified. The procedure, however, presupposes the knowledge of the strata sizes, that is, the total number of sampling units in each stratum as well as the availability of a frame for selecting a sample from each stratum.

[10] An exception is discussed in section 4 of chapter 9.

7.2 NOTATION

We use the same notation as for simple random sampling, except that there will be a subscript to indicate a particular stratum when we refer to information regarding this stratum. Thus, N will represent the total number of elements in the population, as before; but N_1 will be the number in the first stratum, N_2 will be the number in the second stratum, etc. Similarly, n will be the total sample size; n_1 will be the size of the sample in the first stratum, n_2 will be the size of the sample in the second stratum, etc. The subscript h denotes the stratum and i the unit within the stratum. As in the case of simple random sampling, capital letters refer to population values and lower case letters denote corresponding sample values. The following notation given in the table will be used.

Measurement	For Population	For Sample	Sample Estimate
Total number of elements	N	n	--
Number of strata	L	L	--
Number of elements in the hth stratum	N_h	n_h	--
Total for a certain variable (characteristic)	Y	y	$\hat{Y}_{st}$
Total of the variable in the hth stratum	Y_h	y_h	$\hat{Y}_h$
Average over all strata (population mean)	$\bar{Y}$	$\bar{y}$	$\bar{y}_{st}$
Average for hth stratum (stratum mean)	$\bar{Y}_h$	$\bar{y}_h$	--
Proportion having attribute	P	p	p_{st}
Proportion in the hth stratum	P_h	p_h	--
Population Variance	S^2	--	--
Variance for the hth stratum	S^2_h	s^2_h	--
Variance of an estimated total	$S^2(\hat{Y}_{st})$	$s^2(\hat{Y}_h)$	$s^2(\hat{Y}_{st})$
Variance of an estimated mean	$S^2(\bar{y})$	$s^2(\bar{y}_h)$	$s^2(\bar{y}_{st})$

Value of a specific unit	Y_{hi}	y_{hi}	--

7.2.1 Illustration for a Whole Population

Suppose we have a universe of eight farms with known value of land and buildings as follows:

Farm	Value of land and buildings
A	$2026
B	6854
C	1532
D	2180
E	5408
F	9284
G	1438
H	8836

Let us compute the average (mean) and the standard deviation of these values. In terms of the notation above, we would have

$$N = 8$$

$$\bar{Y} = \$4,694.75$$

$$S = \$3,326.04$$

Now let us arrange the farms into two strata, so that the groupings of values are as follows:

Stratum 1	Stratum 2
$1,438	$5,408
1,532	6,854
2,026	8,836
2,180	9,284

If we compute the average and standard deviation of each group of four farms separately, we would have

Stratum 1	Stratum 2
$N_1 = 4$	$N_2 = 4$
$\bar{Y}_1 = \$1,794$	$\bar{Y}_2 = \$7,595.50$
$S_1 = \$364.33$	$S_2 = \$1,800.45$

7.3 ESTIMATES FROM A STRATIFIED SAMPLE

The population mean can be expressed in terms of the stratum totals, as follows:

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L} Y_h = \frac{Y}{N} \tag{7.1}$$

where the population total Y = $\sum_{h=1}^{L} Y_h$.

Since each Y_h can be expressed as $N_h\bar{Y}_h$, we may write

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{Y}_h \tag{7.2}$$

Within each stratum, simple random sampling is used. We saw previously that for simple random sampling, $\bar{y}$ is an unbiased estimate of $\bar{Y}$. This suggests that for stratified sampling an estimate of the population mean can be obtained by substituting, for each stratum mean, the corresponding estimate from the sample. That is, the mean of the sample elements from the first stratum gives an estimate of the true mean of the first stratum; the mean of the sample elements in the second stratum gives us an estimate of the true mean for the second stratum, etc. In symbols, therefore, the estimate of the population mean from a stratified sample is denoted by $\bar{y}_{st}$ (st for stratified) and is given by :

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{y}_h \tag{7.3}$$

Another way of expressing the same formula is

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} \frac{N_h}{n_h} y_h \tag{7.4}$$

where y_h is the sample total for the h^{th} stratum.

7.3.1 Illustration of estimate of mean

A stratified sample is drawn from a population of 1,000 farms to estimate average expenditure by farm operators for hired labor. There are three strata--the total number of farms in the first is 300; in the second, also 300; and in the third, 400. The selected samples have 30, 30, and 40 farms in the three strata respectively. The average expenditure for the 30 farms in the first stratum is \$12.20; for the 30 farms in the second stratum, \$25.60; and for the 40 farms in the third stratum, \$48.70. For the sample estimate of the average expenditure for all farms in the population we would have

$$\bar{y}_{st} = \frac{N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3}{N}$$

$$= \frac{300(12.20) + 300(25.60) + 400(48.70)}{1{,}000}$$

$$= \frac{3{,}660 + 7{,}680 + 19{,}480}{1{,}000} = \$30.82.$$

7.3.2 Estimate of total

As with simple random sampling, we make an estimate of the population total by multiplying the estimate of the mean by the total number of elements in the population:

$$(7.5) \qquad \hat{Y}_{st} = N\bar{y}_{st} = \sum_{h=1}^{L} N_h\bar{y}_h = \sum_{h=1}^{L} \frac{N_h}{n_h} y_h$$

7.3.3 Estimate of proportion

To estimate a proportion for the population, the procedure is similar to that for the mean because a proportion, P_{st} is simply a special case of the mean $\bar{Y}$ when the only possible values oY_i are 0 and 1. In this case,

$$\bar{Y}_h = P_h = \frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi} \quad where \quad Y_{hi} = \qquad 0 \text{ or } 1$$

for stratified random sampling. The population proportion P_{st} is

$$P_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h P_h$$

The estimate of this is

(7.6) $$p_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h p_h \quad where\ p_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$$

7.4 STANDARD ERROR OF A STRATIFIED SAMPLE

The standard errors of the three types of estimates referred to above are computed by using equation (7.7) for the mean, equation (7.8) for the total, and equation (7.9) for a proportion:

(7.7) $$S(\bar{y}_{st}) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h}}$$

where $$S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(Y_{hi} - \bar{Y}_h)^2$$

(7.8) $$S(\hat{Y}_{st}) = NS(\bar{y}_{st})$$

(7.9) $$S(p_{st}) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{P_h Q_h}{n_h}}$$

The corresponding formulas for the estimated standard error for each type of estimate are:

(7.10) $$s(\bar{y}_{st}) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h}}$$

where the sample standard error ,

$$s_h = \sqrt{\sum_{i=1}^{n_h}\frac{(y_{hi} - \bar{y}_h)^2}{n_h - 1}}$$

(7.11) $$s(\hat{Y}_{st}) = Ns(\bar{y}_{st})$$

(7.12) $$s(p_{st}) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{p_h q_h}{n_h}}$$

Similar formulas can be derived for the coefficient of variation (CV) by dividing the above expressions by the value of the item being estimated. Thus, for example:

(7.13) $$cv(\bar{y}_{st}) = \frac{s(\bar{y}_{st})}{\bar{y}_{st}}$$

The formulas for confidence intervals of the population mean and the population total are:

(7.14) $$\bar{y}_{st} \pm ts(\bar{y}_{st})$$

(7.15) $$\hat{Y}_{st} \pm ts(\hat{Y}_{st}) = N\bar{y}_{st} \pm Ns(\bar{y}_{st})$$

These formulas assume that $\bar{y}_{st}$ is normally distributed and that $s(\bar{y}_{st})$ is well determined, so that the multiplier t can be read from tables of the normal distribution (see Appendix I). If only a few degrees of freedom (less than 30) are provided by each stratum the t-value should be taken from the tables of student's t (see an Appendix II) instead of the normal table.

7.4.1 Illustration

Let us apply equation (7.7) to the case of the eight farms in the illustration in section 2. Suppose we took a sample of four farms out of the eight--two from each stratum--and we have computed $\bar{y}_{st}$ by equation (7.3). What is the standard error of $\bar{y}_{st}$?

In the two strata, the values would be:

Stratum 1	Stratum 2
$N_1 = 4$	$N_2 = 4$
$n_1 = 2$	$n_2 = 2$
$S_1 = 364.33$	S2 = 1,800.45
$S^2 = 132{,}736.35$	$S^2 = 3{,}241{,}620.2$

Applying equation (7.7)

$$S(\bar{y}_{st}) = \sqrt{\frac{1}{N^2}[N_1(N_1-n_1)\frac{S_1^2}{n_1} + N_2(N_2-n_2)\frac{S_2^2}{n_2}]}$$

$$S(\bar{y}_{st}) = \sqrt{\frac{1}{64}[4(4-2)(\frac{132{,}736.35}{2}) + 4(4-2)(\frac{3{,}241{,}620.2}{2})]}$$

$$= \sqrt{\frac{1}{64}(530{,}945.40 + 12{,}966{,}481)}$$

$$= \sqrt{210{,}897.28} = \$459.24$$

It is interesting to compare this standard error with the corresponding error of the mean of a simple random sample of two farms. For a simple random sample of two farms, we would have

$$S(\bar{y}) = \sqrt{(\frac{N-n}{N})\frac{S^2}{n}}$$

$$= \sqrt{(\frac{8-4}{8})\frac{(3{,}326.04)^2}{4}} = \$1{,}175.93$$

In this example, the standard error of the stratified sample is much smaller than that of the simple random sample, less than half. In fact, it would require a sample of six farms, using simple random sampling, to achieve the same reliability (that is, as small a standard error) as we obtained with a stratified sample of the four farms.

7.4.2 Remarks

In actual practice, we usually do not know the true values S_h^2 and P_hQ_h. Instead, we substitute sample estimates of these values into equations (7.7), (7.8), and (7.9) to obtain equations (7.10), (7.11) and (7.12), respectively. To make such estimates from a single sample, we would need at least two elements from each stratum. (In the examples described above, we were able to compute the standard error for samples having only one element per stratum because we had information on all elements in the universe.)

To derive equation (7.7), we do the following:

(7.16) $$\bar{y}_{st} = \frac{N_1\bar{y}_1 + N_2\bar{y}_2 + ... + N_L\bar{y}_L}{N}$$

Apply the variance operator S^2 to each side of equation (7.16).

(7.17) $$S^2(\bar{y}_{st}) = \frac{1}{N^2} S^2[N_1\bar{y}_1 + N_2\bar{y}_2 + ... + N_L\bar{y}_L]$$

(7.18) $$S^2(\bar{y}_{st}) = \frac{1}{N^2} [S^2(N_1\bar{y}_1) + S^2(N_2\bar{y}_2) + ... + S^2(N_L\bar{y}_L)]$$

(7.19) $$S^2(\bar{y}_{st}) = \frac{1}{N^2} [N_1^2 S^2(\bar{y}_1) + N_2^2 S^2(\bar{y}_2) + ... + N_L^2 S^2(\bar{y}_L)]$$

(7.20) $$S^2(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \; S^2(\bar{y}_h)$$

But $S^2(\bar{y}_h) = \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$ for h = 1, 2, 3, ..., L. Therefore, we can write (7.20) as:

$$S^2(\bar{y}_{st}) = \frac{1}{N^2}\left[\sum_{h=1}^{L} N_h^2 \; \frac{N_h - n_h}{N_h} \; \frac{S_h^2}{n_h}\right] \tag{7.21}$$

Equation (7.21) is equivalent to equation (7.7).

We will now rewrite equation (7.21) in a different way to make some observations.

(7.22) $$S^2(\bar{y}_{st}) = \frac{1}{N^2}\left[\sum_{h=1}^{L} \frac{N_h - n_h}{N_h} \frac{(N_h S_h)^2}{n_h}\right]$$

which can also be written the following way:

(7.23) $$S^2(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{(N_h S_h)^2}{n_h} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2$$

From equation (7.22) we can see that if the fpc = 1, i.e., if $[(N_h - n_h) / N_h] = 1$ then e
(7.22) becomes:

(7.24) $$S^2(\bar{y}_{st}) = \frac{1}{N^2}\left[\sum_{h=1}^{L} \frac{(N_h S_h)^2}{n_h}\right]$$

Equation (7.23) has two components. The first component is shown in equation (7.24) and it represents the variance of the mean when sampling with replacement, that is, when the fpc = 1.

The second term in equation (7.23) represents the adjustment that one needs to make when sampling without replacement.

We can also see from equation (7.24) that the variance of the mean is directly proportional to the strata population variance. That is, the smaller the population variance in the strata, the smaller the variance of the mean. In other words, the more homogeneous the strata, the smaller the overall variance of the mean with stratified sampling.

Study Assignment

Problem A: *Suppose you have a population of 12 persons whose hourly earnings are as follows:*

A - $ .85	*E - $1.80*	*I - $1.75*
B - $1.35	*F - $3.10*	*J - $ .75*
C - $ .60	*G - $ .90*	*K - $2.40*
D - $2.20	*H - $1.50*	*L - $2.10*

Exercise 1. *What is the average (mean) hourly earnings for this group?*

Exercise 2. *What is the standard error of the mean for a sample of four persons selected as a simple random sample?*

Exercise 3. *Stratify this population into three strata of equal size in the best way to estimate average earnings. List the persons in each stratum by their hourly earnings.*

Exercise 4. *Select a sample of six persons--two from each stratum:*

(a) Show the formula you would use for estimating the average (mean) hourly earnings for this sample.

(b) Show the formula for the standard error of the estimated mean.

Problem B: *Refer to the universe of eight farms (in section 7.2.1 on page 86 of the text) with known value of land and buildings as follows:*

A - $2026	*E - $5408*
B - $6854	*F - $9284*
C - $1532	*G - $1438*
D - $2180	*H - $8836*

Exercise 5. *Identify the 28 combinations of possible samples of two farms each in simple random sampling (AB, AC, etc.).*

(a) Compute the mean for each sample and verify that the average mean of all 28 means is $4,694.75.

(b) Compute the standard deviation of the 28 means and check that the standard deviation is $2,037.

Exercise 6. *Identify the 36 combinations of possible samples of two farms each in stratified sampling, using the two strata (stratum 1 and stratum 2) in section 7.2.1 of the text.*

(a) Check that the mean of the 36 means is $4,694.75.

(b) Check that the standard deviation of the 36 means is $795.40.

Chapter 8

STRATIFIED SAMPLING-ALLOCATION TO STRATA

8.1 THE PROBLEM OF ALLOCATION

The definition of stratified sampling does not specify a particular size of sample in a stratum. The sample can be selected so as to have the same size in each stratum, or it can be distributed in some other way. As long as we select at least one element per stratum, the specification for a stratified sample is satisfied; and with two elements per stratum we can estimate both the mean and its error. Usually the total sample size is much larger than two elements per stratum. Hence, the question arises as to what criterion should be used in allocating the total sample among the strata.

Let us return to the earlier example of a population of eight farms in two strata. If we wish to select a sample of two farms to estimate the mean, we have no choice but to take one farm from each stratum. Suppose, however, that we wish to select four farms. Then we have a choice in the allocation of the sample. Would it be better to select two farms from each stratum or take one farm from one stratum and three farms from the other?

There are two important criteria for determining how the sample should be distributed among the various strata. The first criterion is convenience; that is, choose a method which is easy to apply and simple to tabulate. This usually leads to the use of proportionate or proportional (allocation) stratified sampling. The second criterion is precision: choose a method which will provide the smallest sampling variance (or sampling error). This leads to the use of optimum allocation.

8.2 PROPORTIONATE STRATIFIED SAMPLING

It is very common in stratified sampling to select the same proportion of units in each stratum. With this method, to take a 10-percent sample of a given population, we would take a 10-percent sample from each stratum.

Since the sampling rates in all strata are the same, the number of elements taken for the sample will vary from stratum to stratum, depending on the size of the stratum. Within each stratum, the sample size will be proportionate to the total population of the stratum. We can express this mathematically as follows:

$$n_h = \frac{n}{N}N_h, \text{ or alternatively } n_h = \frac{N_h}{N}n$$

For the population characteristics that we are usually interested in (namely, Y and $\bar{Y}$), we can prepare estimates from a proportionate stratified sample as easily as from a simple random sample--in fact, by using the same formula

(8.1) $$\bar{y}_{st}=\frac{1}{n}\sum_{i=1}^{n} y_i.$$

In this formula, the sum is for all sample elements without regard to strata; since (N_h/n_h) is a constant, and equal to (N/n), equation (7.3) of chapter 7 reduces to this form. We also have

(8.2) $$\hat{Y}_{st}= N\bar{y}_{st}= \frac{N}{n}\sum_{i=1}^{n} y_i$$

where i in equations (8.1) and (8.2) refers to individual observations.

The simple weighting procedure makes proportionate sampling attractive since results are easy to tabulate. Different strata do not have to be tabulated separately. All of the sample data can be added together before application of any factors such as (1/n) or (N/n). A sample which has this feature is self-weighting. That is, in a self-weighting sample, every individual observation has the same probability of selection and, consequently, the same weight. The true standard error of the mean estimated from a proportionate stratified sample is

(8.3) $$S(\bar{y}_{st})=\sqrt{\sum\left(\frac{N_h-n_h}{N_h}\right)\frac{N_h^2}{N^2}\frac{S_h^2}{n_h}}$$

When we substitute $n_h=\frac{N_h}{N}n$ in equation (8.3), this becomes,

(8.4) $$S(\bar{y}_{st})=\sqrt{\left(\frac{N-n}{Nn}\right)\sum\frac{N_hS_h^2}{N}}$$

(8.5) $$S(\hat{Y}_{st})= NS(\bar{y}_{st}) = \sqrt{\left(\frac{N-n}{n}\right)\sum N_hS_h^2}$$

Proportional allocation has many advantages:

1. In order to use this allocation procedure we don't need to know the stratum variances (as

the methods we'll discuss later do).

2. Other methods require us to know the costs of sampling units in the different strata, but not this method.

3. The increase in precision from other more elaborate methods is not very large.

However, we will see later on that when there is a very large variation in the stratum variances, the gain in precision obtained by other methods may outweigh the simplicity of proportional allocation. However, as shown later, this method is widely used in applied sample design.

8.3 OPTIMUM ALLOCATION

Sometimes we have to conduct a survey with a fixed amount of money and we may be faced with the fact that the cost of sampling units in different strata differs widely. For instance, it is a well-known fact that sampling units in rural areas is generally more expensive than urban areas, because the distances are longer and sometimes sampling units are more difficult to find. The term optimum allocation refers to the optimum (the most efficient) way of allocating the total sample (n) to the different strata. The formula is given by:

$$n_h = \frac{\dfrac{N_h S_h}{\sqrt{c_h}}}{\displaystyle\sum_{h=1}^{L} \frac{(N_h S_h)}{\sqrt{c_h}}} \cdot n$$

where c_h is the cost of sampling one unit in stratum h. The above formula is obtained by finding

the values of n_h that will minimize $S^2(\bar{y}_{st})$ subject to the linear constraint $C = \sum_{h=1}^{L} c_h n_h$.

When the costs of sampling in the different strata are the same, the optimum allocation formula is called Neyman allocation, after Jerzy Neyman (1934), who investigated mathematically the question of what distribution of the sample among strata would give the smallest possible sampling error. He found that the answer was to let the sampling rate in each stratum vary according to the amount of variability in the stratum--in other words, to make the sampling rate in a given stratum proportional to the standard deviation in that stratum. The number of elements to be sampled from any stratum, then, would depend not only on the total number of elements in that stratum, but also on the standard deviation of the characteristic to be measured. For Neyman allocation, the number to be selected within a stratum is given by the following formula:

(8.6) $$n_h = n \frac{N_h S_h}{\Sigma N_h S_h}$$

With Neyman allocation, the formula for the variance of the mean (after using (8.6) in formula (8.3)) reduces to

(8.7) $$S^2(\bar{y}_{st}) = \frac{1}{n}(\frac{1}{N}\sum N_h S_h)^2 - \frac{1}{N^2}\sum N_h S_h^2$$

The second term on the right represents the use of the fpc.

As before, the standard error of the total is given by the following formula:

(8.8) $$S(\hat{Y}_{st}) = NS(\bar{y}_{st})$$

For this type of allocation, it is necessary to know the values of S_h in the universe. If these are not known in advance, then they may be estimated within each stratum, by using the methods described in section 1.3 of chapter 5.

Note that in formula (8.6), when the S_h are all equal, Neyman allocation becomes proportionate allocation.

8.3.1 Illustration

Let us compare the standard errors arising from proportionate and optimum allocation in the same survey. In 1942, a census of lumber production was taken in the United States. In 1943, the survey was to be repeated, but on a sample basis. Before selecting the sample, mills were grouped into strata, on the basis of their 1942 production; an analysis of the data produced the information presented in Table 8.1.

Table 8.1

BASIC DATA FOR DETERMINING OPTIMUM ALLOCATION

(Production figures and standard deviations given in thousands of board feet)

Stratum	1942			Standard deviation for 1943 (S_h)*
	Annual Production	Number of mills (N_h)	Average production in stratum	
1	5,000 and over	538	11,029.7	9,000
2	1,000 to 4,999	4,756	1,779.6	1,200
3	Under 1,000	30,964	203.8	300
Total		36,258	571.2	**1,684

*Estimated from 1942 data. **For unstratified sampling.

Now let us select a sample of 1,000 mills. The first question to consider is how to determine the sample size in each stratum, under either proportionate sampling or optimum allocation sampling. The second question to consider is the resulting reliability of the two methods. Let us consider first the matter of the sample size, then the matter of reliability.

8.3.2 Sample Size in Each Stratum

For proportionate allocation, since the sampling rate is 1,000 out of 36,258, this rate is used in each stratum. The sample sizes, therefore, would be:

$$n_1 = \frac{1{,}000}{36{,}258} \quad x\ 538 = 15$$

$$n_2 = \frac{1{,}000}{36{,}258} \quad x\ 4{,}756 = 131$$

$$n_3 = \frac{1{,}000}{36{,}258} \quad x\ 30{,}964 = 854.$$

For optimum allocation, the sample size in each stratum would be determined by the following table.

Table 8.2

SAMPLE SIZE FOR OPTIMUM ALLOCATION

Stratum	Number of mills (N_h)	Standard Deviation (S_h)	N_hS_h	$\frac{N_hS_h}{\sum N_hS_h}$	Number in sample (n_h)*	Sampling rate
1	538	9,000	4,842,000	0.244	244	1/2
2	4,756	1,200	5,707,200	0.288	288	1/16
3	30,964	300	9,289,200	0.468	468	1/66
Total	36,258		19,838,400	1.000	1,000	

$$*n_h = 1{,}000 \text{ x } \frac{N_hS_h}{\sum N_hS_h}$$

8.3.3 Standard Errors

What are the standard errors for these two sample designs? For proportionate allocation, the standard error of the estimate of the mean is given by equation (8.4):

$$S(\bar{y}_{st}) = \sqrt{(\frac{N-n}{Nn})\frac{\sum N_hS_h^2}{N}}$$

For the survey of lumber production,

$$\sum N_hS_h^2 = 538(9{,}000)^2 + 4{,}756(1{,}200)^2 + 30{,}964(300)^2 = 53{,}213{,}400{,}000$$

and

$$S(\bar{y}_{st}) = \sqrt{\frac{36{,}258 - 1{,}000}{36{,}258(1{,}000)} x \frac{53{,}213{,}400{,}000}{36{,}258}}$$

$$= \sqrt{1427} \qquad = 37.8 \text{ (thousand board feet).}$$

For optimum allocation, the corresponding standard error is given by equation (8.7):

$$S(\bar{y}_{st}) = \sqrt{\frac{1}{n}\left(\frac{\sum N_h S_h}{N}\right)^2 - \frac{\sum N_h S_h^2}{N^2}}$$

$$= \sqrt{\frac{1}{1000}\left(\frac{19{,}838{,}400}{36{,}258}\right)^2 - \frac{53{,}213{,}400{,}000}{(36{,}258)^2}}$$

$$= \sqrt{259} = 16.1 \qquad \text{(thousand board feet)}$$

To complete the analysis, one may compare these results with those obtained if we had not stratified the mills, but had taken a simple random sample of 1,000 mills from the universe. In this case, the standard error is given by:

$$S(\bar{y}) = \sqrt{\frac{(N-n)}{n}\,\frac{S^2}{n}} = \sqrt{\frac{36{,}258-1{,}000}{36{,}258}\,\frac{(1{,}684)^2}{1{,}000}} = 52.5 \ (\textit{thousand board feet})$$

8.4 COMPARISON OF SAMPLING ERRORS WITH DIFFERENT SAMPLING METHODS

Examining the results of the sample designs above, we see that optimum allocation gave us a standard error of 16.1 thousand board feet, considerably smaller than that under proportionate sampling, which was 37.8; we see also that the sampling error under proportionate sampling was smaller than that under simple random sampling, which was 52.5. Putting the results another way, it would require a proportionate sample more than 5 times as large as an optimum allocation sample to achieve the same reliability. Simple random sampling would require a sample 10 times as large. The efficiency of optimum allocation results from the fact that it provides for more intensive sampling in strata having large standard deviations, which can be expected to contribute more heavily to the total sampling error.

The example in section 3 above illustrates a general result which can be demonstrated mathematically. The sampling errors of the three types of designs are approximately related in the following way (if the sampling rates are small enough so that the finite correction factors can be ignored):

(8.9) $$S^2_{ran} = S^2_{opt} + \frac{\sum N_h(s_h - \bar{S})^2}{nN} + \frac{\sum N_h(\bar{Y}_h - \bar{Y})^2}{nN}$$

(8.10) $$= S^2_{prop} + \frac{\sum N_h(\bar{Y}_h - \bar{Y})^2}{nN}$$

where $\bar{S} = \frac{\sum N_h S_h}{N}$ is a weighted average of the values of S_h. S^2_{ran}, S^2_{opt}, S^2_{prop} and respectively the variances of the estimated means based on simple random sampling, optimum and proportionate sampling.

An examination of this formula shows that sampling errors obtained with optimum allocation will be at least as small, and usually smaller, than those obtained with proportionate stratified sampling. Furthermore, the errors obtained with either of these methods will be at least as small, and generally smaller, than those obtained with simple random sampling. (There are a few rare cases, which almost never occur in practice, in which this is not true. When the sample is very small and the stratification is completely ineffective, neither proportionate sampling nor optimum allocation may show a gain over simple random sampling. For all practical purposes, this possibility can be ignored.)

Consider the conditions under which important differences result from the three methods. When we compare proportionate stratified sampling with simple random sampling, it can be shown that the gain in reliability depends on the amount by which the means of the strata vary; the greater the variation between the means (in other words, the greater the differences among the strata), the more the reduction in the standard error arising from the use of proportionate sampling. On the other hand, if the variance between stratum means is fairly small compared to the total variance, not much will be gained by stratification. As a result, stratification is usually less important in dealing with proportions than with measured items (or with aggregates or quantities). For example, it would be of much greater help in trying to estimate the average expenditure of farmers for hired labor than in estimating the proportion of farmers who hire labor. Even for measured items the gains would be slight unless the strata are established so that the differences between the means are sizable (as was the case in the example of lumber mills). For example, in conducting a survey to measure personal income, it would probably not pay to establish separate strata for different professional groups--for example, doctors, lawyers, etc. It probably would be useful, however, to set up separate strata for broader groups--laborers, businessmen, professionals, etc. Since proportionate sampling is nearly always better than simple random sampling, stratification is recommended whenever it can be accomplished with little additional work.

Comparing optimum allocation with proportionate allocation, we see that if the standard deviations in all strata are the same, the two methods are identical. The greater the differences between the standard deviations in the strata, the greater the reduction in sampling error to be expected from optimum allocation. Unless the range among the standard deviations is greater than 2 or 3 to 1, the gains of optimum allocation are so small that they are probably not worth the extra complications in tabulation. With larger variations in standard deviations, the gains are appreciable and optimum allocation is advisable. In the example of lumber mills, the standard deviation for stratum 1 was 30 times as large as that for stratum 3.

We need to know the S_h for each stratum either (a) to apply optimum allocation or (b) to estimate the errors of proportionate stratified samples. Of course, in practice, we never really know each S_h and must estimate it. Two questions arise: (a) How is the accuracy of the sample affected by the errors introduced by estimating S_h instead of knowing the true value? (b) What methods can be used to estimate these quantities?

In answer to the first question, if our estimates of the standard deviation are fairly reasonable (for example, accurate to within 30% or 40%) we will obtain almost all of the gains of optimum allocation. The reason for this is that the sampling error does not increase very rapidly as the allocation departs from the optimum within fairly broad limits. (It should be noted that poor guesses of the values of S_h do not introduce any biases in the result; they only increase the sampling errors.) However, if the estimates of S_h are very unreliable, the "optimum allocation" may have a larger variance than proportionate allocation. In this case, it is safer to use proportionate allocation.

In regard to the second question, we can use the methods for estimating the standard deviations described previously (section 1.3 of chapter 5). One additional method that is sometimes used is to assume that the standard deviations for the strata are proportional to the average values within the strata; that is, assume the same relative standard deviation in each stratum. (Note that for optimum allocation, it is not necessary to know the absolute values of the standard deviations; it is only necessary to know their values relative to each other.) This assumption will frequently give results reasonably close to the optimum. In the case of the lumber mills discussed previously, this would give us a sample with the following distribution by strata:

$$n_1 = n\frac{N_1\bar{Y}_1}{\sum N_h\bar{Y}_h} = 1{,}000\ x\ \frac{538(11{,}029.7)}{20{,}708{,}219} = 287$$

$$n_2 = n\frac{N_2\overline{Y}_2}{\sum N_h\overline{Y}_h} = 1{,}000\ x\ \frac{4{,}756(1{,}779.6)}{20{,}708{,}219} = 409$$

$$n_3 = n\frac{N_3\overline{Y}_3}{\sum N_h\overline{Y}_h} = 1{,}000\ x\ \frac{30{,}964(203.8)}{20{,}708{,}219} = 305$$

It can be seen that this allocation is much closer to optimum allocation than is proportionate allocation. In fact, if the standard error of this allocation is computed, it turns out to be 17.3. This is not quite as good as the 16.1 for optimum allocation, but it is far superior to the 37.8 obtained with proportionate sampling.

8.5 OPTIMUM ALLOCATION WITH VARIABLE COSTS

The discussion of optimum allocation thus far has been in terms of getting the most reliable results for a given total sample size. It frequently happens that the costs of obtaining information vary substantially from stratum to stratum. To give an example, let us suppose that families have been stratified by urban and rural residence; furthermore, suppose that the cost of conducting a rural interview is five times as great as that of an urban interview. It would be wise to concentrate more of the sample in the cheaper stratum. Another example would be a sample survey of business firms; we may mail questionnaires to small companies and visit large ones personally, when there are large differences in unit costs.

A more general approach than the one which is described in section 4 above is to consider the optimum allocation for a fixed cost, rather than for a fixed sample size. In other words, we would like to allocate the sample among strata in such a way as to achieve the lowest standard error with a fixed budget.

For this we need a cost function, which is a mathematical formulation expressing the cost of taking the survey in terms of the sample sizes, n_h. Suppose the average cost for a single questionnaire in the hth stratum is called C_h. Thus C_1 is the cost per questionnaire in the first stratum, C_2 is the cost in the second stratum, etc. C_h represents the total cost of a questionnaire in the hth stratum, including the cost of interviewing, coding, data entry, etc. (There may also be an overhead cost for the survey which does not depend on the size of the sample, but it is not necessary to consider this in the cost function.) The total cost of the survey which can be affected by the sample size is

$$C = C_1n_1+C_2n_2+C_3n_3+\ldots+C_Ln_L = \sum_{h=1}^{L} C_hn_h$$

For a fixed cost C, the optimum allocation of the sample turns out to be[11]

(8.11)
$$n_h = n \times \frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum \frac{N_h S_h}{\sqrt{C_h}}}$$

That is, n_h is directly proportional to N_h and to S_h, and inversely proportional to $\sqrt{C_h}$

Formula (8.11) leads to several rules. In a given stratum, we would take a larger sample under the following conditions:

(1) If the stratum is larger than the average stratum.

(2) If the stratum is more variable internally than the average stratum.

(3) If the cost of collection and processing is cheaper than in the average stratum.

In regard to the third point, the cost per stratum (C_h) enters into the formula in the form of a square root. This tends to reduce the effect of the differences in unit cost. Unless the costs vary by a factor of at least 2 to 1, using the formula above will give results not very much different from the simpler optimum allocation given in equation (8.6).

In equation (8.11), we do not yet know the value of n. If cost is fixed, substitute the value of n_h from (8.11) in $C = \sum C_h n_h$ and solve for n. This gives

(8.12)
$$n = \frac{C \sum (N_h S_h / \sqrt{C_h})}{\sum (N_h S_h \sqrt{C_h})}$$

If, however, an estimate with a specified variance, S, is required, n is given by

(8.13)
$$n = \frac{(\sum N_h S_h \sqrt{C_h})(\sum N_h S_h / \sqrt{C_h})}{N^2 S + \sum N_h S_h^2}$$

[1] To use this formula, n must first be calculated. n is a function of C, and the c_h's, S_h's, and N_h's. See Sample Survey Methods and Theory, Volume I: Methods and Applications, by Hansen, M.H., Hurwitz, W.N., and Madow, W.G. New York, Wiley and Sons, 1953, p. 221.

Note that in case $C_h = c$, that is, if the cost per unit is the same in all strata, then the cost becomes $C = \sum cn_h = cn$ and also equation (8.11) reduces to equation (8.6). That is, optimum alloca

for fixed cost reduces to optimum allocation for fixed sample size.

8.5.1 Illustration

Suppose a sampler proposes to take a stratified random sample. He expects that his field costs will be of the form $C = \sum C_h n_h$. His advance estimates of relevant quantities for the two s

are as follows:

Stratum 1	Stratum 2
$N_1 = 1{,}056$	$N_2 = 1{,}584$
$S_1 = 10$	$S_2 = 20$
$C_1 = \$4$	$C_2 = \$9$

(a) Find the sample size required under optimum allocation, to make $S(\bar{y}_{st}) = 1.$ Ignore the fpc.

(b) Determine the sample size for each stratum (i.e., the allocation of the total sample size n to each of the two strata).

(c) How much will the total field cost be (excluding overhead costs)?

Solution of (a)

For optimum allocation, the standard error of the estimate of the mean, ignoring the fpc, is given by equation (8.7):

$$S(\bar{y}_{st}) = \sqrt{\frac{1}{n}\left(\frac{\sum N_h S_h}{N}\right)^2}$$

$$1 = \sqrt{\frac{1}{n}\left(\frac{\sum N_h S_h}{N}\right)^2}$$

Thus,

$$n = (\frac{\sum N_h S_h}{N})^2$$

Given the estimates for each stratum specified above,

$$\sum N_h S_h = (1056)(10)+(1584)(20) = 42240$$

$$N = N_1+N_2 = 1056+1584 = 2640$$

and

$$n = (\frac{42240}{2640})^2 = 256$$

Solution of (b)

The sample size is given by equation (8.11).

$$n_h = n \times \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum \frac{N_h S_h}{\sqrt{c_h}}}$$

The sample size for the first stratum would be:

$$n_1 = n \times \frac{\frac{N_1 S_1}{\sqrt{c_1}}}{\left(\frac{N_1 S_1}{\sqrt{c_1}}+\frac{N_2 S_2}{\sqrt{c_2}}\right)}$$

$$= 256 \times \frac{\frac{(1056)(10)}{2}}{\frac{(1056)(10)}{2}+\frac{(1584)(20)}{3}} = 85$$

Similarly,

$n_2 = 171$

The total field cost is given by:

$C = C_1n_1 + C_2n_2 \ \ = 4\ (48) + 9\ (171) = \$1,875$

8.6 OPTIMUM ALLOCATION FOR SEVERAL ITEMS

The formula for the optimum allocation of the sample (equation 8.6 or 8.11) is necessarily computed for a single characteristic or variable, Y. If it is desired to obtain the most favorable sample allocation for several characteristics, some kind of compromise must be made.[12] Some alternatives are:

(1) Determine the most important item (or group of highly correlated items) and allocate the sample to get the best estimate for this item.

(2) Follow the procedure in (1) and increase the size of the sample in some strata to provide adequate coverage of other important items.

(3) Set up a function which assigns a weight to each item according to its importance; use this function in the allocation to prevent poor sample estimates for the most important characteristics.

Optimum allocation is most effective for characteristics which vary widely for the individual units; such as amount of personal income, number of board feet produced by a sawmill, kilos of maize harvested on a farm, etc.

In sampling for attributes, however, such as the proportion of the population in a class (for example, in the income class \$1,000 - \$1,999), proportionate sampling may be the best allocation. It has the added advantage of being self-weighting.

8.7 STRATIFIED SAMPLING FOR PROPORTIONS

Before concluding this chapter, some comments will be made on the problem of sample allocation when the object is to estimate a population proportion P. From equation (7.9 of chapter 7, we have for stratified random sampling,

$$S(p_{st}) = \sqrt{\frac{1}{N^2}\sum \frac{N_h^2(N_h - n_h)}{N_h - 1}\frac{P_hQ_h}{n_h}} \tag{8.14}$$

[2] For a further discussion of the problem of sampling for many characteristics, see section 13 of chapter 5.

with proportional allocation,

$$(8.15) \qquad S(p_{st}) = \sqrt{\frac{(N-n)}{N}\frac{1}{nN}\sum\frac{N_h^2P_hQ_h}{N_h-1}}$$

Ignoring the fpc,

$$(8.16) \qquad S(p_{st}) = \sqrt{\frac{1}{nN}\sum N_hP_hQ_h(\frac{N_h}{N_h-1})}$$

For the sample estimate of the variance, substitute $\frac{p_hq_h}{(n_h-1)}$ for the unknown $\frac{P_hQ_h}{n_h}$ in a

the formulas above.

If the optimal allocation can be used, n_h will be chosen proportional to $N_h\sqrt{P_h(1-P_h)}$. This

allocation will differ substantially from proportional allocation only if the quantities $\sqrt{P_h(1-P_h)}$ differ considerably from stratum to stratum. For example, let the P_h lie between 0.3 and 0.7, in which case $\sqrt{P_h(1-P_h)}$ will lie between 0.46 and 0.50. In this situation the optimum allocation will not be preferred to proportional allocation when the simplicity of the computations involved is another factor to be taken into account.

We can choose n_h in order to minimize the variance $S(p_{st})$ from section 5.

<u>Minimum variance for fixed total sample size.</u>

$$n_h \propto N_h\sqrt{\frac{N_h}{N_h-1}}\sqrt{P_hQ_h} \doteq N_h\sqrt{P_hQ_h}$$

where $\propto$ represents proportional to

Thus,

(8.17) $$n_h \doteq n \frac{N_h \sqrt{P_h Q_h}}{\sum N_h \sqrt{P_h Q_h}}$$

Minimum variance for fixed cost.

(8.18) $$n_h \doteq n \frac{N_h \sqrt{P_h Q_h / C_h}}{\sum N_h \sqrt{P_h Q_h / C_h}}$$

where cost = C = $\sum C_h n_h$

The value of n is found by substituting $S_h = \sqrt{P_h Q_h}$ in equation (8.12) or (8.13).

8.7.1 Illustration

In a firm, 62% of the employees are skilled or unskilled males, 31% are clerical females, and 7% are supervisory. A sample of 400 employees is taken from a total of 7,000 employees. Based on the sample, the firm wishes to estimate the proportion that uses certain recreational facilities. Rough guesses are that the facilities are used by 40 to 50% of the males, 20 to 30% of the females, and 5 to 10 % of the supervisors. How would you allocate the sample among the three groups ? What would the standard error of the estimated proportion p_{st} be? Ignore the fpc.

We have,

N = 7,000 n = 400
N_1 = 4340, N_2 = 2170 and N_3 = 490

We guess P_1 = 45%, P_2 = 25% and P_3 = 7.5% as a compromise.

Using equation (8.17), we can allocate the total sample size (n = 400) to the different strata, as follows:

$$n_1 = n \times \frac{N_1 \sqrt{P_1}}{\sum N_h \sqrt{P}}$$

$$= 400 \times \frac{(4340)\sqrt{(.45)(.55)}}{(4340)\sqrt{(.45)(.55)} + (2170)\sqrt{(.25)(.75)} + (490)\sqrt{(.075)(.925)}} = 268$$

Similarly,

$n_2 = 116$ and $n_3 = 16$

The standard error is given by the equation (8.16):

$$S(p_{st}) = \sqrt{\frac{1}{nN} \sum N_h P_h Q_h}$$

$$= \sqrt{\frac{(4340)(.45)(.55)+(2170)(.25)(.75)+(490)(.075)(.925)}{(7{,}000)(400)}} = 0.02326$$

8.8 DETERMINATION OF SAMPLE SIZE n

In simple random sampling, we saw that the determination of n depended on the sampling variance of the estimator. In a similar way, for stratified sampling, we need to know the formulas for the sampling variances of the different methods of allocation in order to determine n for each one of these methods. Let's summarize the methods of allocation.

1. Equal samples from each stratum.

2. Proportionate allocation.

3. Optimum allocation: fixed budget, varying sampling costs among strata.

4. Neyman allocation: fixed sample size, equal sampling costs among strata.

We also saw that the stratum sample sizes n_h for these methods of allocation were given by:

(8.19) $$n_h = \frac{n}{L} \qquad (\textit{equal samples})$$

(8.20) $$n_h = \frac{N_h}{N}n \qquad (\textit{proportionate})$$

(8.21) $$n_h = \frac{N_h S_h / \sqrt{c_h}}{\sum (N_h S_h / \sqrt{c_h})} n \qquad (\textit{optimum})$$

(8.22) $$n_h = \frac{N_h S_h}{\sum (N_h S_h)} n \qquad (\textit{Neyman})$$

1. To determine the sample size we need to know the variances of these methods. So, we start with the formula for the variance of a mean when using stratified random sampling. Recall that the formula is given by:

$$S^2(\bar{y}_{st}) = \frac{1}{N^2}\left[\sum_{h=1}^{L} N_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{S_h^2}{n_h}\right] \qquad (8.23)$$

Now we substitute the different values of n_h into formula (8.23).

After we do this, we get:

$$S^2(\bar{y}_{eq}) = \frac{L}{N^2}\sum_{h=1}^{L} \frac{N_h^2\ S_h^2}{n} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2 \qquad (8.24)$$

$$S^2(\bar{y}_{prop}) = \frac{1}{N}\sum_{h=1}^{L} \frac{N_h\ S_h^2}{n} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2 \qquad (8.25)$$

$$S^2(\bar{y}_{opt}) = \frac{1}{N^2}\ \frac{1}{n}\ (\sum_{h=1}^{L} N_h\ S_h\sqrt{c_h}) \left(\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}\right) - \frac{1}{N^2}\sum N_h S_h^2 \qquad (8.26)$$

$$S^2(\bar{y}_{Ney}) = \frac{1}{N^2}\frac{(\sum_{h=1}^{L} N_h\ S_h)^2}{n} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2 \qquad (8.27)$$

Now, let's see how to determine the sample size n to estimate the mean with an error of estimation E. The sample size is directly related to the error we are willing to tolerate (or the precision we are required to obtain) in our estimates. As before, we define the error the following way:

Error of estimation = E = k S($\bar{y}$)

where k is the level of reliability. So, given the precision E that we need to obtain and the level of reliability k, we can write:

(8.28) $$S^2(\bar{y}_{st}) = \frac{E^2}{k^2} = B^2$$

We know that as n increases, the variance of the estimate becomes smaller. Therefore, we need to find the sample size n that will give us a variance equal to B^2.

Let's try to solve for n in equation (8.24), that is, when we have equal samples.

(8.29) $$B^2 = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2\ S_h^2}{n} - \frac{1}{N^2}\sum_{h=1}^{L}N_h S_h^2$$

Multiply each side of equation (8.29) by N^2 and leave the term which contains n on one side of the equation. After we do this, we obtain:

(8.30) $$N^2B^2 + \sum_{h=1}^{L}N_h S_h^2 = \frac{1}{n}L\sum_{h=1}^{L}N_h^2\ S_h^2$$

When we solve for n, we obtain:

(8.31)
$$n = \frac{L\sum_{h=1}^{L} N_h^2 \; S_h^2}{N^2B^2 + \sum_{h=1}^{L} N_h S_h^2}$$

Now, when (n_h/N_h) is very small (negligible), the fpc = 1 and we may omit from the denominator of equation (8.31) the term $\sum N_h S_h^2$.

Applying a similar procedure to equations (8.25), (8.26), and (8.27), we obtain the sample size n given by the following formulas:

(8.32)
$$n_{prop} = \frac{N\sum_{h=1}^{L} N_h \; S_h^2}{N^2B^2 + \sum_{h=1}^{L} N_h S_h^2}$$

(8.33)
$$n_{opt} = \frac{(\sum N_h \; S_h\sqrt{c_h})(\sum N_h \; S_h/\sqrt{c_h})}{N^2B^2 + \sum N_h S_h^2}$$

$$(8.34) \qquad n_{Neyman} = \frac{(\sum N_h \; S_h)^2}{N^2B^2 \; + \; \sum N_h S_h^2}$$

As before, when the fpc = 1, the denominator in equations (8.32), (8.33) and (8.34) only contains the term $N^2B^{2..}$. Another important point to mention is that all the formulas for n have been given in terms of the stratum population variances (S_h). In practice, we don't know this value and it has to be estimated by means of a sample or from other sources.

Study Assignment

Problem A: *It is desired to estimate the total value of farm products for a population of 5,900 farms. Means and variances are available from a past census on the value of farm products classified by farm size and tenure of the operator:*

Size and tenure	*Number of farms* *(N_h)*	*Average value of products* $(\bar{Y}_h)$	*Variance* (S_h^2)	*Standard deviation* (S_h)	$N_h \; S_h$

All farms.........	*5,900*	*$ 3,500*	*97,000,000*	*$ 9,840*	
SIZE OF FARM					
Under 10 acres....	*590*	*1,200*	*18,000,000*	*4,240*	*2,502,000*
10 to 49 acres....	*1,600*	*1,500*	*15,000,000*	*3,870*	*6,192,000*
50 to 99 acres....	*1,150*	*2,200*	*18,000,000*	*4,240*	*4,876,000*
100 to 179 acres..	*1,200*	*3,600*	*35,000,000*	*5,920*	*7,104,000*
180 to 259 acres..	*490*	*5,500*	*70,000,000*	*8,370*	*4,101,000*
260 to 999 acres..	*650*	*6,200*	*200,000,000*	*14,150*	*9,198,000*
1,000 and over....	*220*	*18,000*	*400,000,000*	*20,000*	*4,400,000*
Product sum.......	**349,620,000,000*				*38,373,000*
TENURE OF OPERATOR					
Full owner.......	*3,300*	*2,600*	*35,000,000*	*5,920*	*19,536,000*
Part owner........	*660*	*6,900*	*110,000,000*	*10,490*	*6,923,000*
Manager...........	*50*	*18,000*	*510,000,000*	*22,580*	*1,129,000*
Tenant............	*1,890*	*3,500*	*40,000,000*	*6,320*	*11,945,000*
Product sum.......	**289,200,000,000*				*39,533,000*

**Product sum =* $\sum N_h S_h^2$

Exercise 1. *Compute the standard error of the total value of products from a proportionate stratified sample of 300 farms for each of the two methods of stratification (by size and by tenure of the operator).*

Exercise 2. *Which method of stratification is more efficient for a proportionate sample?*

Exercise 3. *Compute the standard error of the estimate of the total value of products, using a simple random sample of 300 farms.*

Exercise 4. *For both methods of stratification, use the Neyman allocation for a sample of 300 farms, and compute:*

(a) The number of sample farms in each stratum

(b) The standard error of the estimate of the total value of products.

Exercise 5. *On the basis of this analysis, which of the four methods of allocating the sample would you recommend?*

Exercise 6. *Assume that the sample was stratified by tenure and allocated by the optimum method. Assume also that the following means by strata were obtained:*

Tenure	*Mean value of products*
Full owner	*$ 2,900*
Part owner	*6,400*

Manager	*20,000*
Tenant	*4,000*

Estimate the mean value of products for the population of 5,900 farms.

Exercise 7. *Describe how you would calculate the standard error of the mean computed in exercise 6 after the survey results are available?*

Problem B: *The following data show the stratification of all the farms in a county by farm size and the average acres of corn per farm in each stratum.*

Farm Size (acres)	*Number of farms* N_h	*Average Corn Acres* $\bar{Y}_h$	*Standard deviation* S_h
0-40	*394*	*5.4*	*8.3*
41-80	*461*	*16.3*	*13.3*
81-120	*391*	*24.3*	*15.1*
121-160	*334*	*34.5*	*19.8*
161-	*430*	*52.0*	*28.6*
Total or mean	*2,010*	*26.3*	

Exercise 8. *For a sample of 100 farms, compute the sample size in each stratum under*

(a) Proportional allocation

(b) Optimum allocation

(c) Simple random sampling

Exercise 9. *For a sample of 100 farms, compute the standard error of the estimated total for*

(a) a simple random sample

(b) a proportional allocation

(c) an optimum allocation

Exercise 10. *On the basis of this analysis, which of the three methods of allocating the sample would you recommend?*

Problem C: *In illustration 5.1, suppose that after the sample is taken, the sampler finds that his field costs were actually $3.00 per unit in stratum 1 and $10.00 in stratum 2.*

Exercise 11. *How much greater is the total field cost than anticipated?*

Problem D: *With three strata, the values of the N_h, S_h, and c_h are as follows:*

Stratum	N_h	S_h	c_h
1	*860*	*5*	*2*
2	*640*	*4*	*3*
3	*1230*	*6*	*5*

Exercise 12. *Find the sample size in each stratum for a sample of size 200 under an optimum allocation.*

Exercise 13. *How much will the total field cost be?*

Problem E: *In illustration 7.1, suppose that the true proportions of users were 48, 21, and 4%, respectively.*

Exercise 14. *How would you allocate the sample among the three groups?*

Exercise 15. *What would the standard error of the estimated proportion* p_{st} *be?*

Exercise 16. *What would the standard error of p be with a simple random sample with n = 400?*

Problem F: *Using the list of 600 households residing in 30 villages (Appendix IV), select a SRS-WOR of 20 households, and on the basis of the data on the size of these 20 sample households, do the following :*

Exercise 17. *Determine the number of households for each zone and then select a sample of size n_h(h = 1, 2, 3) in each of the three zones. Use proportional allocation.*

Exersise 18. *Estimate for each of the 3 zones separately:*

(a) the total number of persons and its standard error.

(b) the average household (HH) size and its standard error.

Exercise 19. *Estimate for the entire population :*

(a) the total number of persons and its standard error.

(b) the average HH size and its standard error.

(c) The coefficient of variations (CVs) for both number of persons and average HH size.

Exercise 20. *Compare the population estimates and standard errors obtained from exercise 19 with those obtained from SRS-WOR in chapter 5.*

Chapter 9

CLUSTER SAMPLING

9.1 DESCRIPTION OF CLUSTER SAMPLING

The discussion so far has been about sampling methods in which the units of analysis (people, farms, business firms, etc.) were considered as arranged in a list (or its equivalent) and a sample of individual units could be selected directly from the list. Now we will consider a sampling procedure in which the units of analysis in the population are grouped into clusters and a sample of clusters (rather than a sample of individual units of analysis) is selected. The sample clusters then determine the units to be included. The determination may be made in either of two ways:

(1) The sample could include all units in the selected clusters. This is usually referred to as single-stage cluster sampling.

(2) A subsample of units in the selected clusters could be selected for enumeration. This is called multi-stage cluster sampling, or simply multi-stage sampling.

There are two main reasons for using cluster sampling. Often there is no adequate frame (such as a list) from which to select a sample of the elements in the population, and the cost of constructing such a frame may be too great. In other cases, such a frame may exist but the savings in field costs obtained by cluster sampling (on some kind of geographical basis) may make this method more efficient than a simple random sample from a list. In most practical situations, a sample of a given number of units selected at random will have smaller variance than a sample of the same size selected in clusters; nevertheless, when cost is balanced against precision, the cluster sample may be more efficient.

Even though the units in which we are interested are not selected directly, the probability of selecting a cluster and each unit in it (i.e., the probability of selecting a unit from the population) is fixed in advance; consequently, cluster sampling satisfies the criterion for probability sampling.

Let us consider some examples to see how cluster sampling works.

9.1.1 Single-Stage Cluster Sampling

To draw a sample of persons, it would generally not be feasible to obtain a list of all persons, and then to select a sample from the list. It might be possible to find a list of families. We could then select a sample of families and obtain information by interview concerning all persons in the selected families. This is an example of single-stage cluster sampling; the family constitutes the cluster. Note that for a given number of individuals in the sample, it would undoubtedly be less costly in terms of both travel and time to take all persons within selected families than to select the same number of persons at random from all individuals in the population.

Often there is no list of families available, and some other procedure must be used. A possible method is as follows. In large cities, a map showing the boundaries of city blocks can usually be obtained; and we can select a sample of blocks. In the rest of the country, we can use maps divided into small areas called segments, which have identifiable boundaries, and select a sample of segments. Within the sample blocks and segments, we could include all persons in the sample; alternatively, we could select a sample of persons living in the selected blocks. The choice would depend upon the number of stages of sampling we believe would be most efficient. By using maps, we eliminate the need for a list of all persons. We replace it with a list of blocks and segments and a list of families within a sample of blocks and segments. (In practice there frequently is an earlier stage of sampling in which a sample of cities and/or other administrative areas is selected.) The preceding discussion illustrates an important application of cluster sampling; namely, area sampling. However, other applications of cluster sampling are frequently made.

9.1.2 Multi-Stage Cluster Sampling

Suppose we wish to make a survey of school children in order to obtain information on their health, or information on their knowledge of a particular subject. One way to do this is to obtain a complete list of schools, then select a sample of schools, and finally choose a sample of children within the selected schools. Similarly, a sample of factory workers could be selected by first choosing a sample of factories and then interviewing a sample of workers within these factories. In both cases we would need to construct a list of individuals only for the schools or factories selected in the sample. These examples illustrate multi-stage (specifically, two-stage) cluster sampling. The probability that any unit in the population is selected in the sample can be expressed as the product of the probabilities at each stage. Thus, in the first example the probability of selecting the j^{th} child from the i^{th} school is the probability of first selecting the i^{th} school times the conditional probability of selecting the j^{th} child, given that the i^{th} school has been selected. That is,

$$P(j^{th}\text{ child, }i^{th}\text{ school}) = P(i^{th}\text{ school}) \text{ x } P(j^{th}\text{ child } \mid \; i^{th}\text{ school}).$$

9.2 AREA SAMPLING

Since area sampling is a frequently used application of cluster sampling, we shall describe in more detail the methods which are usually applied. Area sampling is useful when one or both of the following conditions exist:

(1) When complete lists of housing units (or other desired units of observation) are not available but maps having a reasonable amount of detail are available. Such maps can be considered as a list covering all of the housing units in the area.

(2) When there are large travel costs in sending an interviewer from one randomly selected sample housing unit to another randomly selected housing unit. For a given amount of money, we may be able to increase the number of sample housing units greatly by grouping units together and selecting a random sample of groups.

Three simple procedures exist for drawing an area sample. We shall use city blocks as an illustration (segments of land with identifiable boundaries around them could be used in rural areas in exactly the same way as blocks are used in cities). We shall assume that a 1-percent sample of housing units is to be drawn.

Procedure A for a sample of areas to be enumerated completely:

(1) Obtain a reasonably accurate map of the city, showing as much detail as possible for blocks. If the map is not new, one should take steps through local inquiry to bring it up-to-date (for example, draw in new streets that have been opened since the map was printed).

(2) Number the blocks serially, entering the numbers directly on the map; a serpentine numbering system is advisable in order to make certain that no blocks are omitted.

(3) Select a simple random or systematic sample of blocks, using a 1-percent sample. If a systematic sample is used, select a random number from 1 to 100 to determine the first sample block, and include every one-hundredth block thereafter.

(4) Interview all households in the sample blocks.

Procedure B for a sample of areas with subsampling of smaller areas:

The 1-percent sample can also be obtained by drawing, for example, a sample of 1 in 25 blocks, then taking a subsample of one-fourth of the area in each sample block.

(1) Proceed as in (1), (2), and (3) in procedure A above, except that instead of taking 1 in 100 blocks, take 1 block in 25.

(2) Divide each of the sample blocks into 4 segments. If maps are available that show the internal structure of each block (alleys, buildings, etc.), these can be used. If not, make a quick and crude sketch of the sample blocks, showing each building; use this sketch as the basis of the segmentation. The 4 segments within any block should have roughly the same number of housing units in each.

(3) Number the segments in each block from 1 to 4.

(4) Select the sample segments by taking a random number from 1 to 4 for each block.

(5) Interview all households in the selected segments.

Notice that although a 1-percent sample is obtained in both procedures, procedure B includes more sample blocks and fewer housing units per block. Usually, it will cost more to obtain the same sample size by procedure B, since there is a cost of subsampling not involved in procedure A; also, travel will be increased in visiting a greater number of blocks. This subsampling procedure is almost equivalent to dividing every block in the city into 4 parts, or segments, and taking 1 in 100 of these segments. Hence, the use of subsampling as described above in procedure B can be regarded as essentially equivalent to using a sample of small clusters of housing units (in which every housing unit would be enumerated) but with two-stage sampling as a device for reducing the work of drawing a sample of small clusters.

Procedure C for a sample of areas with listing and subsampling:

To carry out procedure B, it is necessary to have or to construct detailed maps. A third procedure accomplishes approximately the same results and is frequently applicable when detailed maps are not available and are not easy to prepare.

(1) Proceed as in step (1) of procedure B, again selecting a sample of 1 in 25 blocks.

(2) Visit each sample block and make a list of all the housing units in it. Number the housing units serially. The numbering can be done (a) separately by blocks (that is, starting with 1 for each block), (b) in a single sequence throughout all the sample blocks, or (c) by some combination, such as a separate sequence for various groups of blocks.

(3) Select one-fourth of the housing units within the sample blocks either by using a random number table, or by systematic sampling using the serial numbers assigned to the housing units.

(4) Interview the households whose serial numbers are selected for the sample.

Note: If advance information is available on the approximate numbers of housing units in all blocks, some combination of the above procedures with stratification of blocks by size can be used.

9.3 CHOICE OF SAMPLING UNIT AND SAMPLE DESIGN

In designing a sample, the sampling statistician must decide how many sampling stages are to be used. In addition, at each stage he must determine the sampling unit. In making his decision, the statistician often has many alternatives from which to choose. Suppose, for example, that he desires to estimate the average number of cattle per holding. Ultimately, the information must be obtained from a sample of individual holdings (units of analysis or elementary units). In order to obtain such a sample, however, any of the following plans could be used:

(1) A simple random, systematic, or stratified sample of individual holdings could be taken if complete and accurate lists of holdings were available.

(2) Maps could be used to subdivide the country into small area segments (for example, segments containing an average of 5 or 10 holdings). A sample of these area segments could then be selected, and all holdings within each selected segment included in the sample. For holdings which extend across segment boundaries, rules would be needed to associate holdings with segments.[13]

(3) A sample of small administrative subdivisions, such as districts, could be selected. All holdings in the selected districts could be included in the sample, or a subsample of holdings could be selected.

(4) A sample of provinces (larger administrative divisions) could be selected, and a sample of areas and holdings within the selected provinces could be taken in one of the ways described in procedures A, B, and C above.

Where subsampling is used, the cluster initially selected is called the first-stage unit or the primary sampling unit (PSU) and the unit of subsampling is called the second-stage unit (SSU). For example, in (3) above, if a subsample of holdings is selected, the "district" is the PSU and the holding is the second-stage unit; in (4), the "province" is the PSU, the small area is the second-stage unit, and still smaller areas or holdings may be third-stage units (TSU).

How can one make an intelligent choice among the various alternatives? We may reason as follows: where cost is not important, single-stage sampling using the elementary unit (the holding in the above case) as the sampling unit provides the most accurate results <u>for the given number of elementary units</u> in the sample. (There are some exceptions, but these are rather unusual cases.) On the other hand, when cost and administrative convenience are important, a cluster sample involving one or more stages may be desirable. The cost of enumeration per elementary unit is usually much less if the units are in clusters than if they are randomly distributed throughout the country; by clustering, travel time and cost for interviewing are reduced. As a result, for a given amount of money it may be possible, by using cluster sampling, to increase the number of elementary units in the sample above the number that the same budget would allow if these were

[1] For further discussion, see section 14 of chapter 12 of Sample Survey Methods and Theory (referred to in footnote 1 of chapter 8).

selected at random. If the increase in the number of units more than compensates for the fact that a cluster sample tends to increase the standard error, a net gain will be obtained in the reliability of estimates made from the sample.

In order to choose among alternative sampling units, we must therefore balance the expected costs against the standard errors for the various possible designs and use the method which will provide the smallest standard error for a fixed cost. In some administrative situations, the correct decision may be obvious. If the survey involves little or no travel cost--for example, if mail questionnaires are used, or if the survey uses personnel who travel around as a normal part of their other activities, such as policemen or postmen (mailmen)--and if listings of elementary units are available, the elementary unit should always be taken as the sampling unit. If travel costs or the costs of constructing lists of elementary units are rather large, an alternative design using a clustered sample will usually be better. A full discussion of this matter is beyond the scope of these chapters, but some of the important points will be discussed here.

9.4 ANALYSIS OF COSTS

Usually there is a fixed budget available for a survey, and one of the major functions of the sampling statistician is to provide a method of obtaining the smallest sampling error for this budget. Let us first examine how costs enter into a survey involving the use of cluster sampling.

In studying stratified sampling, we discussed the possibility that enumeration and processing costs can vary from stratum to stratum, and we constructed a cost function which expressed the variable part of the total cost as a sum of unit costs multiplied by sample sizes (for example, $C = C_1 n_1 + C_2 n_2 + ...$). A similar approach is needed for cluster sampling, although the unit costs are of a different type. For simplicity, let us consider a two-stage sample.

9.4.1 Components of Cost

In order to analyze the costs of a two-stage cluster sample, it is necessary to identify the various phases of the survey and to distinguish between three elements of cost:

(1) Overhead costs; that is, those costs that are fixed regardless of the manner in which the sample is selected.

(2) Costs that depend primarily on the number of first-stage clusters in the sample, and the way in which such costs vary as the number of these primary sampling units in the sample varies.

(3) The costs that depend primarily on the number of second-stage units in the sample, and the way in which such costs vary with this number.

9.4.1.1 Overhead Costs

Overhead costs include such things as the administrative and technical work required for the survey, rent for space and for some types of equipment, cost of printing the final results, etc. These costs will generally be approximately the same, even with great variations in the size and design of the survey. Since these costs are not affected by the size of the survey, they do not enter into the decision on sample design. The only reason for separating these costs is to subtract them from the total available budget in order to see what funds can be spent on the variable costs.

9.4.1.2 Costs of First-Stage Units

Certain costs will usually vary in proportion to the number of first-stage sampling units. These will include (a) the cost of selecting, traveling to, and locating each first-stage unit, (b) the cost of preparing a list of second-stage units (within the primary unit), and (c) the cost of designating the subsample of second-stage units. There may also be other costs (costs of preparing maps for the first-stage sample units, hiring special enumerators to handle each one, etc.) depending on the nature of the administrative organization, and the materials available before the start of the survey.

9.4.1.3 Costs of Second-Stage Units

The costs which depend on the number of second-stage units will include the costs of interviewing, reviewing the survey results, coding, recording, etc.

9.4.2 A Simple Cost Function

Let us assume a simple situation in which the cost per first-stage unit does not change despite changes in the number of such units in the sample. Similarly, the cost per second-stage unit does not change. Then the total variable cost (which excludes overhead costs) can be represented by

$$C = C_1 m + C_2 n = C_1 m + C_2 m\bar{n} \tag{9.1}$$

where

C_1 is the cost per first-stage unit,
C_2 is the cost per second-stage unit,
m is the total number of first-stage sampling units.
n is the total number of second-stage sampling units.
$\bar{n}$ is the average number of second-stage units in a primary unit.

Using equation (9.1), one can set down combinations of m and n which would add up to the same cost. For example, suppose the total variable cost available for a survey was \$2,500, and the estimates of C_1 and C_2 were \$10 and \$2, respectively. The table below shows various combinations of sample sizes all of which would cost exactly $2,500; the last column shows the average size of cluster $\bar{n}$ for each allocation:

Number of units in sample		Average
First stage (m)	Second stage (n)	$(\bar{n} = \frac{n}{m})$
10	1200	120
20	1150	57.5
50	1000	20
75	875	11.7
100	750	7.5
125	625	5
150	500	3.3

If the standard error can be found for each of the above combinations, one can choose that combination which would give the lowest standard error. In fact, with this simple type of cost function, it is usually possible to determine the optimum allocation mathematically. However, this is not necessary; if a formula can be found which expresses the variance in terms of m and n, we can easily see which combination is best. Furthermore, this can also be done in situations involving more complex cost functions, when it is more difficult to develop a mathematical solution to the problem of optimum allocation. The next chapter will be devoted to analyzing the variances for the simpler and more common situations.

9.4.3 More Complex Cost Functions

One additional comment on costs should be made. The formulation of the cost function above as $C = C_1 m + C_2 n$ covers the simplest type of situation only. In practice, the cost function may be much more complex. For example, there may be stratification for either the first-stage or the second-stage units with different unit costs in each stratum. The cost function would then be

$$C = \sum C_{1i}\, m_i + \sum C_{2i}\, n_i$$

and the problem of the allocation of the sample would be a combination of optimum allocation for cluster sampling with optimum allocation for stratified sampling. Frequently, the unit costs would depend on the number of units in the sample.

For example, suppose that C_1 included a part that resulted from the time spent traveling from one first-stage unit to another. With only a few primary units in the sample, the average distance from one to the next might be quite large, resulting in a high value of C_1. However, as the number of units in the sample increases, the average distance gets smaller and C_1 will be smaller. A different type of cost function would be used in such a situation. In general, in planning a large-scale and important survey, a detailed analysis should be made of how costs vary, in order to construct a cost function which is realistic for that particular survey.

STUDY ASSIGNMENT

Problem A: *On the following page is a map of a district which is divided into six subdistricts. Each subdistrict is divided into a varying number of EA's (Enumeration Areas). Suppose you wished to select a 12 ½-percent (1 out of 8) sample of holdings from this district.*

Exercise 1. *Describe how you would select a 12 ½-percent sample of holdings using a 25-percent sample of EA's and a 50-percent sample of holdings within the selected EA's (using an approach similar to procedure C in lecture 9 of the text for selecting holdings within the sample EA's).*

Exercise 2. *Describe how you would select a 12 ½-percent sample of holdings using a 12 ½-percent sample of EA's.*

Exercise 3. *Suppose you first selected a 50-percent sample of subdistricts. How might you proceed from there to select you 12 ½-percent sample of holdings in a total of two stages? In a total of three stages?*

Problem B: *A survey is to be conducted to provide information on production of a certain crop which can be grown only under a government-controlled permit program. Permits issued at the start of the planting season will be used as a source of information; these permits are issued by district agricultural offices. The sample will be a two-stage sample: first, a sample of district agricultural offices will be selected by a sampling technician. Enumerators will then visit the selected offices, list the permit holders, and select a sample of holders. The enumerators will then visit the sample holdings to gather the production data. Because the district offices are widely scattered, a different enumerator will be required to handle each one.*

Exercise 4. *Listed below are some of the items of cost for the survey. Indicate with a check mark in the appropriate column below whether the cost should be primarily considered part of the overhead cost, first-stage unit cost, or second-stage unit cost.*

		First	*Second*
	Overhead	*Stage*	*Stage*
a. Printing data collection forms..........			
b. Training enumerators....................			
c. Obtaining a list of permit offices......			
d. Visiting permit offices to select sample of permit holders.............			
e. Selecting sample of permit holders......			
f. Collecting the data from the holdings...			
g. Supervisor's field check on the enumerator's work....................			
h. Editing data collection forms...........			
i. Preparing a program to tabulate the survey results...................			
j. Preparing the final report..............			

Problem C: *Refer to the table in section 4.2 of the text for lecture 9.*

Exercise 5. *Verify that the cost for the various combinations of sample sizes totals $2,500; assume that the cost is $10 per first-stage unit and $2 per second-stage unit.*

CHAPTER 10

CLUSTER SAMPLING VARIANCES

10.1 VARIANCE OF A TWO-STAGE CLUSTER SAMPLE

To study the variance of a two-stage cluster sample, it will be useful to review some ideas of stratified sampling. In stratified sampling, the standard error of a sample estimate depends on the within-stratum variances, S_i^2. For each stratum, the varian(S_i^2) is defined by the same formula as S^2 (the total variance of the population) but using only the elements in the i^{th} stratum. We saw that stratified sampling was most useful when the means of the strata were very different. In fact the gains of stratified sampling can be determined by computing the standard deviation among the means of the strata (that is, computing the standard deviation of the numbers $\bar{Y}_1, \bar{Y}_2, \bar{Y}_3,$ *etc.*, weighted by the number of units within each stratum) if the necessary data are available. The square of this weighted standard deviation between stratum means is called the between-strata variance.

Similar concepts can be considered in cluster sampling. In fact, there is a close analogy between cluster and stratified sampling. In both cases we group the individual elements into sets before selecting the sample. The difference is that in stratified sampling it is necessary to sample within every one of the sets (the strata); in cluster sampling a sample of the sets (the clusters) is selected and then either all or a sample of the elements within the selected sets is included. The purpose and method of forming the sets is very different in the two cases.

10.1.1 Notation

Consider a two-stage design in which second-stage sample units (SSUs) are selected randomly from the elementary units within selected clusters (primary sampling units or PSUs) for interview.

N = total number of PSUs (first-stage clusters) in the population.

n = number of selected sample PSUs.

$M = \sum_{i=1}^{N} M_i$ = total number of elementary units (second-stage units) in the population

$m = \sum_{i=1}^{n} m_i$ = total number of second-stage units (SSUs) in the sample

M_i = number of SSUs in the i-th PSU (cluster), where i = 1,..., N

$\bar{M} = \frac{1}{N}\sum_{i=1}^{N} M_i = \frac{M}{N}$ = average number of SSUs per PSU in the population or ave

cluster

size.

m_i = number of SSUs selected for the sample in the ith PSU, i = 1,..., n

$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ = average number of SSUs per sample PSU

Y_{ij} = value of a characteristic for the jth elementary unit in the ith PSU in the population

$Y_i = \sum_{j=1}^{M_i} Y_{ij}$ = total value of the characteristic in the ith PSU in the population

$Y = \sum_{i=1}^{N} Y_i$ = total value of the characteristic in the population

$\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} Y_{ij}$ = average value of the population characteristic in the ith PSU

$\bar{Y}_c = \frac{1}{N}\sum_{i=1}^{N} Y_i$ = average value of the characteristic per PSU (cluster) in the population

$\mu = \bar{Y} = \frac{Y}{M}$ = Population mean per unit.

y_{ij} = value of the characteristic for the jth sample SSU in the ith sample PSU

$y_i = \sum_{j=1}^{m_i} y_{ij}$ = total value of characteristic in the ith sample PSU

$\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$ = sample average of the characteristic in the ith sample PSU

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y}_c)^2$$ = Population variance between PSU totals

$$S_{wi}^2 = \frac{1}{M_i-1}\sum_{j=1}^{M_i}(Y_{ij}-\bar{Y}_i)^2$$ = within PSU variance in the ith PSU (for populatio

10.1.2 Estimates of Means and Totals

The formulas given in previous chapters for estimating population means are appropriate when the sampling unit is identical with the unit of analysis. An important characteristic of cluster sampling, however, is that the sampling unit (at least in the first stage) is not the unit of analysis. Thus, in the examples in the previous chapter, we would probably not be interested in the mean per family, per school, per factory, or per block. Rather, we would be interested in estimating the mean per family member, per school child, per factory worker, or per housing unit.

Consider a two-stage design in which the second stage units are the units of analysis; n clusters are selected from among N clusters by simple random sampling; and m_i units are selected in the ith PSU using simple random sampling for i = 1, ... , n.

Within the ith cluster, the population mean per unit is given by

(10.1) $$\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i}Y_{ij}$$

Since the units within the cluster were selected by simple random sampling, we know (from chapter 4, section 2) that we can estimate this mean without bias, by using the following formula:

(10.2) $$\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i}y_{ij}$$

These estimates of the cluster unit means from the n sample clusters must then be combined in some way to estimate the overall population total (Y) and the population mean per unit $\left(\mu = \bar{Y} = \frac{Y}{M}\right)$. Several estimators are available and are discussed in most standard texts: we shall examine only one of these.

First, we shall construct an estimator for the population total for the Y-characteristic. An unbiased estimator for Y_i, the ith PSU total is given by

$$(10.3) \qquad \hat{Y}_i = \frac{M_i}{m_i} y_i$$

An unbiased estimator for the population total is then given by

$$(10.4) \qquad \hat{Y} = \frac{N}{n}\sum_{i=1}^{n}\hat{Y}_i = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}y_i = \frac{N}{n}\sum_{i=1}^{n}M_i\bar{y}_i = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}y_{ij}$$

Similarly, we can estimate the total number of units of analysis in the population (assuming that we do not know it) by

$$(10.5) \qquad \hat{M} = \frac{N}{n}\sum_{i=1}^{n}M_i$$

The population mean per unit is

$$\mu = \bar{Y} = \frac{Y}{M} = \frac{\sum_{i=1}^{N}Y_i}{\sum_{i=1}^{N}M_i}$$

An estimator of $\bar{Y}$ is

$$(10.6) \qquad \hat{\mu} = \bar{y} = \frac{\hat{Y}}{\hat{M}} = \frac{N}{\hat{M}}\frac{\sum_{i=1}^{n}M_i\bar{y}_i}{n} = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{\hat{M}}\sum_{j=1}^{m_i}\frac{y_{ij}}{m_i}$$

As can be seen, this estimator is a weighted mean of the n sample cluster means per unit where the weights are the corresponding cluster sizes. As indicated previously, this is only one of several possible estimators; however, this estimator seems to be most generally useful. Since both the numerator and denominator are random variables, this is a ratio-type estimator and it has the usual bias of a ratio estimator. The bias will usually not be serious if the number of clusters in the sample is reasonably large. Ratio estimators are discussed in more detail in chapter 11.

10.1.3 Variances

Consider the case when n PSUs are selected from a population of N PSUs and random samples of m_i (i = 1,...,n) SSUs are taken from the M_i (i=1,...,N) SSUs in the selected PSUs. Then the variance of $\hat{Y}$, the estimator of Y is

(10.7) $$S^2(\hat{Y}) = N^2 \frac{(N-n)}{N} \frac{S_b^2}{n} + \frac{N}{n} \sum_{i=1}^{N} M_i^2 \frac{(M_i - m_i)}{M_i} \frac{S_{wi}^2}{m_i}$$

where,

(10.8) $$S_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y}_c)^2$$ and

(10.9) $$S_{wi}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2$$

The variance of the estimator of Y is the sum of two components. The first component is the contribution to the variance arising from the selection of first-stage units. The second component is the contribution from the selection of second-stage units. If there are three or more stages of sampling, the variance will include additional terms similar in form for each additional stage.

The sample estimator of $Var(\hat{Y})$ is

$$(10.10) \qquad s^2(\hat{Y}) = N^2\frac{(N-n)}{N}\frac{s_b^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i^2\frac{(M_i-m_i)}{M_i}\frac{s_{wi}^2}{m_i}$$

unbiased for

$S^2(\hat{Y})$ although s_b^2 is not unbiased for S_b^2 Now,

$$(10.11) \qquad s_b^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(M_i\bar{y}_i - \frac{1}{n}\sum_{i=1}^{n}M_i\bar{y}_i)^2$$

and,

$$(10.12) \qquad s_{wi}^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2$$

The variance of $\bar{y} = \hat{\mu}$, the estimator of $\bar{Y} = \mu$, is more complex. It is given approxi
by:

(10.13)

$$S^2(\bar{y}) = S^2(\hat{\mu}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{n}\right)\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}\right)^2(\bar{Y}_i-\bar{Y})^2 + \left(\frac{1}{\bar{M}}\right)^2\left(\frac{1}{Nn}\right)\sum_{i=1}^{N}\frac{M_i^2(M_i-m_i)}{M_im_i}S_{wi}^2$$

The approximate value of the variance of $\bar{y}$ may also be obtained from equation (10.7) as follows:

$$(10.14) \quad S^2(\bar{y}) = S^2\left(\frac{\hat{Y}}{M}\right) = \frac{1}{M^2}S^2(\hat{Y}) = \frac{1}{M^2}\left[\frac{N^2(N-n)}{Nn}S_b^2 + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(M_i-m_i)}{M_im_i}S_{wi}^2\right]$$

and is estimated using

(10.15) $$s^2(\bar{y}) = \frac{1}{M^2}\left[\frac{N^2(N-n)}{Nn}s_b^2 + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i^2(M_i-m_i)}{M_i m_i}s_{wi}^2\right]$$

If all PSUs have the same number of second-stage units M and a constant number m of them is sampled from every sample PSU, we have

$$M_i = \bar{M} = \frac{M}{N}, \quad m_i = \bar{m} = \frac{m}{n}$$

and (10.16) $$\hat{Y} = \frac{N}{n}\frac{\bar{M}}{\bar{m}}\sum_{i=1}^{n}\sum_{j=1}^{\bar{m}} y_{ij}$$

In this case, the variance of $\hat{Y}$ is

(10.17) $$S^2(\hat{Y}) = N^2\frac{(1-f_1)}{n}S_b^2 + N^2\bar{M}^2\frac{(1-f_2)}{\bar{m}n}S_w^2$$

where, $f_1 = \frac{n}{N}$, $f_2 = \frac{\bar{m}}{\bar{M}}$, and

(10.18) $$S_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_{wi}^2$$

The sample estimate of $S^2(\hat{Y})$ is given by:

(10.19) $$s^2(\hat{Y}) = N^2\frac{(1-f_1)}{n}s_b^2 + N^2\bar{M}^2\frac{(1-f_2)}{\bar{m}n}s_w^2$$

where, (10.20) $$s_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_{wi}^2$$

The variance of an estimated mean is

(10.21) $$S^2(\bar{y}) = \frac{V(\hat{Y})}{M^2} = \frac{1}{M^2}\left[N^2\frac{(1-f_1)}{n}S_b^2 + N^2M^2\frac{(1-f_2)}{mn}S_w^2\right]$$

and is estimated using

(10.22) $$s^2(\bar{y}) = \frac{1}{N^2\bar{M}^2}\left[N^2\frac{(1-f_1)}{n}s_b^2 + N^2M^2\frac{(1-f_2)}{mn}s_w^2\right]$$

10.1.3.1 Illustration

A population consists of four clusters of five households each. The second-stage units, which are also the elementary units in this case, are houses having persons as follows:

	Cluster			
Household	1	2	3	4
1	3	8	4	7
2	10	3	6	2
3	9	6	3	6
4	8	4	8	4
5	6	5	6	6

First, select two clusters at random from a population of four clusters. Then within each of these selected clusters take a random sample of three households. Compute $\hat{Y}$, the estimate of the population total Y and the variance of $\hat{Y}$. Find the variance and the coefficient of variation of $\bar{y}$, the estimate $\bar{Y}$ of

In this case, N = 4, n = 2, M$_i$= $\bar{M}$ = 5, and m$_i$ $\bar{m}$ = 3

Suppose that clusters 3 and 4 are selected at random. Assume also that households 1, 2, and 5 within cluster 4 and households 2, 4, and 5 within cluster 3 are selected at random. Then we have,

$$\hat{Y} = \frac{N\bar{M}}{n\bar{m}} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} = \frac{(4)(5)}{(2)(3)}[7+2+6+6+8+6] = 116.67 \ \textit{persons}$$

Using equation (10.19),

$$s^2(\hat{Y}) = N^2 \frac{(1-f_1)}{n} s_b^2 + N^2 \bar{M}^2 \frac{(1-f_2)}{\bar{m}n} \sum_{i=1}^{n} \frac{s_{wi}^2}{n}$$

where $s_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (M_i\bar{y}_i - \frac{1}{n} \sum_{i=1}^{n} M_i\bar{y}_i)^2$ $\quad s_{wi}^2 = \frac{1}{m-1} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2$, and

We have,

$$\bar{y}_i = \frac{1}{m} \sum_{j=1}^{m} y_{ij}, \quad \bar{y}_1 = \frac{7+2+6}{3} = 5, \quad \bar{y}_2 = \frac{6+8+6}{3} = 6.67$$

and $\frac{1}{n} \sum_{i=1}^{n} M_i\bar{y}_i = \frac{1}{2}[25+33.33] = 29.17$

$$s_{w1}^2 = \frac{1}{3-1} \ [(7-5)^2 + (2-5)^2 + (6-5)^2] = 7,$$

$$s_{w2}^2 = \frac{1}{3-1} \ [(6-6.67)^2 + (8-6.67)^2 + (6-6.67)^2] = 1.33$$

$$\sum \frac{s_{wi}^2}{n} = \frac{7+1.33}{2} = 4.17$$

On substitution, we have,

$$s^2(\hat{Y}) = 250.51$$

The average number of persons per household is estimated by:

$$\bar{y} = \frac{\hat{Y}}{N\bar{M}} = \frac{116.67}{(4)(5)} = 5.83 \textit{ persons}$$

The estimated variance of this estimate is:

$$s^2(\bar{y}) = \frac{s^2(\hat{Y})}{N^2\bar{M}^2} = \frac{250.51}{(16)(25)} = 0.63$$

The standard error of $\bar{y}$ is:

$$s(\bar{y}) = \sqrt{0.63} = 0.79$$

and the coefficient of variation of $\bar{y}$ is:

$$\frac{s((\bar{y})}{\bar{y}} = \frac{0.79}{5.83} = 0.14 = 14\%$$

10.1.4 Random Group Method of Approximating Variances

The above formulas are somewhat cumbersome. Consequently, short-cut approximations are often used to reduce the amount of work, particularly if variance estimates are to be computed for a large number of characteristics. One of these approximations is known as the random group method.

The random group method consists of dividing the sample into a number of groups at random; each group is then used to make an estimate of the total, mean, etc. (this would be done for each characteristic for which a variance is to be computed). Each of the random groups will reflect the various steps of the sample selection so that the estimate from each group is an estimate of the total with the same sample design as the whole sample (but with a much smaller sample size). In a multi-stage sample, the random groups are usually formed by placing the entire sample from a primary sampling unit in a single group. For complex designs using stratification and/or sampling over time, somewhat different methods are available to divide the sample into random groups. However, the method is not very useful if the number of first-stage units is small.

In computing the estimates of variance, it is exactly the variance between different possible estimates of the total or mean in which we are interested. Therefore, this method which provides a number of different estimates of the total or mean, each with some degree of stability (that is, the number of cases in a group should not be too small) is a realistic one for estimating variances.[14]

10.2 LIMITING FORMS OF VARIANCE OF TWO-STAGE SAMPLE

Examining the variance equation (10.7) and equation (10.9), we can easily see what happens in two simple situations. First, if all second-stage units are included in the sample we have the case described in chapter 9 as "single-stage cluster sampling." In this case, $m_i = M_i$ and the term arising from variation within first-stage units is zero. In equation (10.7), the first term is the same as the variance formula for simple random sampling except that the sample sizes and values of Y_i refer to the first-stage units. For example, if area segments were the first-stage units, N is the total number of area segments and Y_i is the segment total for the variable. In equation (10.9), the first term is the between cluster component of the overall variance which is based on the differences among cluster means per unit of analysis rather than on differences among cluster totals.

Secondly, consider a situation in which all first-stage units are in the sample. In this case, n = N and the first term becomes zero. The variance of the estimator of the population total becomes equal to

$$\sum_{i=1}^{N} M_i^2 \frac{(M_i-m_i)}{M_i m_i} S_{wi}^2$$

The variance of the estimator of the population mean per element is then equal to

$$\left(\frac{1}{N\bar{M}}\right)^2 \sum_{i=1}^{N} M_i^2 \frac{(M_i-m_i)}{M_i m_i} S_{wi}^2$$

These are the variance formulas for the estimators of totals and means from a stratified sample. In other words, a stratified sample is simply a special case of a cluster sample in which all first-stage units are included in the sample and a subsample of second-stage units is selected from each first-stage unit.

This discussion has covered only the case of simple random sampling for both the first-stage and second-stage selections. Analogous formulas can be developed for stratified cluster sampling in which the only difference is that the terms in the equations are replaced by the sums of similar terms over strata.

[1] Refer to section 16 of chapter 10 of Sample Survey Methods and Theory (referred to in footnote 1 of chapter 8).

10.3 ANALYSIS OF COMPONENTS OF VARIANCE

A more detailed analysis of equation (10.7) and equation (10.13) would show that for a two-stage sample containing a given total number of units of analysis, the sampling variances of estimates computed by equation (10.4) and equation (10.6) depend on several factors. Two important factors which the sampling statistician must consider in designing the sample are:

(1) The variability in size of first-stage units in terms of the number of second-stage units they contain.

(2) The variability among second-stage units (the elementary units or units of analysis) within first-stage units.

10.3.1 Variability in Size of First-Stage Units

If the first-stage units are unequal in size in terms of the number of second-stage units (for example, the number of holdings in an area segment), these variations in size can have a profound effect on the size of the variance of the estimator of the population total, as shown by the first term in equation (10.7). We can see in equation (10.13) that the variance of the estimator of the population mean per elementary unit is affected by the variation among first-stage means per element. If the variability in size is very great, it will be necessary to use a large sample of first-stage units or to change the sampling and estimating methods to keep the standard error within reasonable bounds (see section 10.4 below).

10.3.2 Variability Among Second-Stage Units

The second important factor is the variability among second-stage units (units of analysis) within first-stage units (clusters). For a given sampling plan in which we select n out of N clusters and an average of $\bar{m}$ units of analysis out of each sample cluster, it can be shown that the greater the variability among second-stage units within first-stage units, the smaller will be the sampling variability of resulting estimates. **In other words, it is desirable that the units of analysis have a relatively low intraclass correlation**. Intraclass correlation is a measure of similarity among units within a cluster with regard to the characteristics being investigated.

A mathematical demonstration of this phenomenon is beyond the scope of this chapter; however, by

considering an extreme example we can gain an intuitive understanding of it. Consider a situation in which the units of analysis within each cluster are identical. Clearly, a sampling plan such as described above would not be efficient. A single unit of analysis within a given cluster would provide complete information about all the units; consequently, the remaining $(\bar{m} - 1)$ units would contribute nothing additional to our knowledge. To include them in the sample would be a waste of resources. The inefficiency of this design in this situation would be reflected in a high sampling variability relative to a simple random sample with the same number of units of analysis.

The statistician must consider the effect of intraclass correlation on the sampling variability when designing a sample. This is particularly true of area sampling since units which are close together geographically are usually quite similar for many characteristics such as income, education, attitudes, type of agricultural activity, etc. The usual approach is to limit the number of units of analysis taken from the first-stage units and include more of the first-stage units in the sample. In a single-stage sample, the statistician can do this by making the clusters as small as practicable. The more common approach, however, is to introduce additional stages in the sampling procedure so that the number of units of analysis ultimately selected from each unit at the last stage is small. The statistician must, of course, balance precision against cost in deciding on a sampling plan.

Notice that in cluster sampling we gain by having units within clusters as unlike as possible, but in stratified sampling we gain by having units within strata as much alike as possible. The reason for this difference becomes clear when you recall from section 10.2 above that in stratified sampling, the "between-cluster" component of the variance drops out of the equation entirely.

10.4 CONTROL OF VARIABILITY IN SIZE OF CLUSTER

In all of this discussion, it has been assumed that the only way we could affect the sampling variance, with the given population, is to take more or fewer sample cases in the first or second stages or to vary the size of the first-stage units. Of course, if the sampling variance can be reduced by appropriate stratification, this should be done first. Several special procedures are also available to control the effect of variability in size of cluster. The most important procedure is described below.

Although this discussion is related to a two-stage sample, a similar analysis could be made for three or more stages. The procedures described below for controlling variability in size apply equally well to first, second, or other stages, whenever cluster sampling is used.

10.4.1 Define Clusters of Equal Size

One obvious method is to attempt to define clusters in such a way that they are approximately equal in size in terms of the number of units of analysis with the expectation that this will tend to make them equal also in terms of characteristics being investigated. If this can be done with available materials and information, then no other action is necessary. For example, if block counts of numbers of housing units are available for cities and villages, it may be possible to group small blocks together to make clusters which contain approximately the same number of housing units.

In some cases, it may be feasible to define clusters directly in terms of a characteristic being investigated. For example, in an agricultural survey, clusters can be constructed to be nearly equal in area. If recent aerial photographs are available, they might even be made nearly equal in terms of cultivated area.

10.4.2 Stratify Clusters by Size

If information is available on the size of all the first-stage clusters in the universe in advance of the survey (reasonably good approximations are adequate), it is possible to stratify the clusters by size group. The effect of stratification is to replace a total variance by a sum of within-stratum variances. Within each stratum, the clusters should be about equal in size; therefore, stratification by size of cluster will have about the same effect as making all clusters in the whole population about equal in size.

If information on size is not available, it may be worthwhile to spend a small amount of the available resources, for example, in making a "Quick Count" of city blocks in order to obtain approximate sizes of the first-stage units (in terms of the number of housing units they contain). Errors in counts do not cause biases in the estimates, which are based on the actual numbers of housing units found in the survey itself.

Either optimum or proportionate sampling can be performed depending on which appears most useful in the particular case. If more than one characteristic is being estimated, proportionate sampling may be preferable to optimum allocation, since the optimum allocation might be different for each characteristic. Also, proportionate sampling is usually safer unless very good measures of size are available, since the use of the optimum allocation formula with poor measures of size may actually increase the variance.

10.4.3 Use of Ratio Estimates

A third method of reducing the effect of variability in cluster size is through the use of ratio estimates. Ratio estimates will be discussed in somewhat greater detail in Chapter 11; an example of the method is given here. A ratio estimate makes use of a quantity of the form $\hat{Y}/\hat{X}$ where both $\hat{Y}$ and $\hat{X}$ are estimates of totals made from sample data. X, the universe total of the quantity of which $\hat{X}$ is an estimate, must be known.[15] One can make a ratio estimate of the universe total Y--an estimate which is frequently very efficient--by using

$$\hat{Y}_R = \frac{\hat{Y}}{\hat{X}} X$$

instead of $\hat{Y}$ alone. The new estimate of $\hat{Y}_R$ thus differs significantly from $\hat{Y}$ since it involves two items having sampling variances instead of one. Ratio estimates are generally much less sensitive to variation in size of cluster than estimates of the type

$$\hat{Y} = \frac{M}{m} \sum_{i=1}^{m} y_i$$

and their use will frequently reduce the standard errors appreciably.

[2] It may be a projection or other figure which is believed to be very close to the true value.

10.4.3.1 Ratio to Approximate Number of Units of Analysis

Two different uses of ratio estimates for this purpose will be discussed. In the first, "X" is a variable closely related to the total number of units of analysis in the clusters and $\hat{X}$ is an estimate, based on the sample clusters only, of the population aggregate, X. For example, consider a sample design in which city blocks are the first-stage units, and housing units are both the second-stage units and the units of analysis. We have rough counts (X_i) of housing units for each block based on a previous census or special counts made for this purpose. These counts can be totaled for all blocks in the city to obtain X. Then $\hat{X}$, a sample estimate of X, can be obtained by adding up the rough counts for the sample blocks only, and multiplying this by N/n (where N is the total number of blocks in the city and n is the number in the sample). Then, a ratio estimate of Y is

$$\hat{Y}_R = \frac{\hat{Y}}{\hat{X}}X$$

If subsampling is used within the first-stage units, the procedure would be modified. In order to make the fullest gain with this type of ratio estimate, it is advisable not to subsample independently within the clusters, but to treat the second-stage units within the clusters as a continuous list and sample systematically throughout.

10.4.3.2 Ratio to a Correlated Statistic

In a second use of ratio estimates the true value of some universe total X is known and a sample estimate $\hat{X}$ (of X) can be obtained in the survey. If the characteristics "Y" and "X" are positively correlated, then $\frac{\hat{Y}}{\hat{X}}X$ will reduce the effect of variability in cluster size (and possibly other types of variability as well). For example, suppose a survey is planned to measure the total wage and salary earnings of factory workers (Y). We can do this by taking a sample of factories (the clusters) and including all workers within the sample of factories. Suppose the total sales of all factories can be found (X) from some other source--tax records, for example. We could then include on our questionnaire to the sample factories a question on total sales (x_i) as well as wage and salary payments (y_i), and we could prepare estimates of population totals for both characteristics from the sample in the usual manner. The ratio estimate of wages and salaries would then be $\frac{\hat{Y}}{\hat{X}}X$.

10.4.4 Use of Probability Proportionate to Size

A fourth method for controlling the effects of variability in cluster size is to select the sample clusters with probability proportionate to size instead of using a simple random sample of clusters. Probability proportionate to size is frequently abbreviated as PPS. Selection with PPS means that a cluster which is, for example, 5 times as large as another, will have 5 times the chance to be in the sample. It might appear, at first, that this would introduce a bias in the sample result, with some clusters overrepresented and others underrepresented. When PPS is used, the unbiased estimate of the total, where there is no subsampling, is

$$\hat{Y} = \sum_{i=1}^{n} \frac{Y_i}{P_i}$$

Here Y_i is the total in the ith cluster in the sample and P_i is the probability of selection of this cluster. It can easily be shown that this provides an unbiased estimate of Y.

10.4.4.1 Two-stage Sampling

A common application of sampling with PPS is the use of PPS for the selection of the first-stage units in a two-stage sample. When this is done, the subsampling rates are usually set as inversely proportional to size. As a result, the chance of any second-stage unit being included in the sample is the product of the probability of the first-stage and second-stage selections. All second-stage units therefore have identical probabilities and the sample is self-weighting.

There are a number of other advantages to this type of selection procedure; for example, the workload can be made approximately the same for all selected first-stage units. Moreover, the estimates will have smaller variances than those from a proportionate sample in which the first-stage units are selected with equal probabilities.

10.4.4.2 Measures of Size

In order to select with PPS, it is necessary to have measures of size of each cluster in the population. If measures of size are not available, it will usually be found worth the effort to prepare crude estimates of size. (Rough approximations will be almost as effective as more exact measures.) Let us assume such measures are available. The mechanics for selecting a sample with PPS can best be described through an illustration.

10.4.4.3 Illustration

Suppose the clusters are blocks and we wish to sample the housing units in a universe made up of the 10 blocks as listed in column (1) of Table 10.1. We would list, in column (2), the measure of size for each block (this may be a rough estimate of the number of housing units), and cumulate the measures of size in column (3). The last figure in column (3) is the total number (rough estimate) of housing units in all 10 blocks. Let us assume that we wish to include in the sample 5 blocks out of the 10, and that the sample is to include 10 percent of all the housing units.

Table 10.1 SELECTION OF SAMPLE BLOCKS

Block numbe r (PSU) (1)	Measur e of size (2)	Cumulativ e Measure (3)	Sample desig- nation (4)	Probability of selection $(P_i) = n_h*(M_{hi}/M_h)$ (5)	Within Cluster Sampling Rate m_{hi}/M_{hi} (6)
1	50	0 - 50	22.5	50 ÷ 60.2	60.2 ÷ 500
2	12	51 - 62			
3	20	63 - 82			
4	31	83 - 113	82.7	31 ÷ 60.2	60.2 ÷ 310
5	10	114 - 123			
6	60	124 - 183	142.9	60 ÷ 60.2	60.2 ÷ 600
7	55	184 - 238	203.1	55 ÷ 60.2	60.2 ÷ 550
8	13	239 - 251			
9	30	252 - 281	263.3	30 ÷ 60.2	60.2 ÷ 300
10	20	282 - 301			

After completing the first three columns of Table 10.1 as shown, proceed as follows:

(1) Since there are 5 blocks in the sample, divide the final cumulative measure (301) by 5; this gives 60.2, which is the "sampling interval" for selecting blocks.

(2) Choose a random number between 00.1 and 60.2; suppose the number happens to be 22.5. This number is called the Random Start (RS).

(3) Use this random number as the starting number and enter it in column (4), on the line whose cumulative measure interval includes the number 22.5. In our example, the cumulative measure interval is [0 - 50].

(4) Add the sampling interval (60.2) to the random start (22.5), that is add 60.2 to 22.5. This number is equal to 82.7; enter 82.7 on the line whose cumulative measure interval contains this number. In our case, the interval is [83 - 113]. Continue adding 60.2 to the last number obtained (82.7 in our case) and obtain the next one: 142.9. Locate the interval which contains 142.9. In our case the interval is [124 - 183]. Continue with this procedure until a number is reached which is larger than the last cumulative measure.

(5) The blocks with entries in column (4) are the ones in the sample. In this example, they are blocks 1, 4, 6, 7, and 9.

(6) The probability (P_i) of selection of each block actually selected is entered in column (5). For each block, the probability is the measure of size in column (2) divided by the sampling interval 60.2.

(7) The sampling rate to be used within each selected block is computed and entered in column (6). For each block, the rate is the desired overall probability of selection, namely 1/10, divided by the entry in column (5). Thus, for block 1, the rate in column (6) would be

$$\text{o}\frac{1}{10} \div \frac{50}{60.2} = \frac{60.2}{500} \qquad 60.2 \div 500 \qquad \text{r}$$

(8) It occasionally happens that some of the blocks are so large that the measures of size are greater than the sampling interval. As a result, there may be two or more entries in column (4) for the same block. In such a case, the subsampling rate within the block is adjusted to make the overall probability for the selection of housing units equal to 1/10, in our example.

STUDY ASSIGNMENT

Problem A: *A population consists of four clusters. The second-stage units, which are also the elementary units in this case, are houses having rental values as follows:*

	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*
	$100	*$100*	*$10*	*$50*
	100	*100*	*20*	*90*
	200		*40*	
	400		*50*	
TOTALS	*800*	*200*	*120*	*140*

Exercise 1. *What is the value of* S_b^2 *(the between-cluster variance)?*

Exercise 2. *What is the value of the within-cluster variance for the first cluster?*

Exercise 3. *A sample of two clusters is selected with equal probability; within each selected cluster, half the elementary units are in the sample.*

(a) How would you compute $\hat{Y}$ *the estimate of Y?*

(b) What is the variance of the sample estimate of the total $S^2(\hat{Y})$?

© *What is the probability that any elementary unit will be in the sample?*

(d) Compute the coefficient of variation for the estimate.

Problem B: *The following table shows areas of cacao holdings of 15 farmers in five clusters (PSUs) of equal size. The five clusters were selected at random from a total of 40 clusters into which the territory had been divided. Each PSU represents a geographic division containing 120 cacao farmers.*

Area of Cacao Holdings

	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
	96	*110*	*102*	*140*	*132*
	134	*121*	*113*	*142*	*162*
	152	*146*	*157*	*161*	*184*
TOTALS	*382*	*377*	*372*	*443*	*478*

Exercise 4. *Estimate the total area of cacao for the territory.*

Exercise 5. *Estimate the average area of cacao per farm.*

Exercise 6. *Compute the standard errors for the estimates given in exercises (4) and (5).*

Exercise 7. *Compute the coefficient of variation for the estimates given in exercises (4) and (5).*

Problem C: *Assume a city with 12 blocks, as listed in the first column below. Measures of size (approximate number of housing units in each block) are given in the second column. On the basis of this information, we wish to select a sample of 4 blocks with probability proportionate to size, and then to select housing units within the blocks in order to obtain a self-weighting sample of an expected 10 housing units.*

Block Number (PSU)	*Approximate Number of Housing Units (measure of size)*	*Cumulative Measure*	*Actual Number of Housing Units**	*Serial Numbers of Actual Housing Units*
1	*10*	*10*	*9*	*1 to 9*
2	*5*	*15*	*6*	*10 to 15*
3	*2*	*17*	*2*	*16 to 17*
4	*5*	*22*	*6*	*18 to 23*
5	*5*	*27*	*6*	*24 to 29*
6	*10*	*37*	*8*	*30 to 37*
7	*10*	*47*	*8*	*38 to 45*
8	*2*	*49*	*2*	*46 to 47*
9	*2*	*51*	*4*	*48 to 51*
10	*5*	*56*	*6*	*52 to 57*
11	*5*	*61*	*6*	*58 to 63*
12	*10*	*71*	*9*	*64 to 72*
TOTALS	*71*		*72*	

**The number of housing units that would actually be found in the block in a field operation if the block were selected in the sample.*

***Exercise 8**.* *Prepare a worksheet similar to table 10 (in chapter 10) and select the sample of blocks. Assume 3.7 is the random start number for designating the sample blocks.*

***Exercise 9**.* *Assume that you have visited the blocks selected in your sample and determine the actual number of housing units as given in the fourth column above. The housing units that actually exist in each block are designated by "Serial Numbers" as shown in the fifth column. Perform necessary computations for selecng the sample of housing units and list the Serial Numbers for the housing units selected in your sample.*

Problem D: *Consider an Appendix IV which contains a list of 600 households of 30 villages located in 3 zones. Using a two-stage cluster sample design, it is desired to estimate the total number of persons in the population. A random sample of four clusters is chosen and five households in each sampled cluster are randomly selected. Assume* $M_i = \bar{M} = 20$ *households and consider the village as the cluster (PSU) for the survey.*

Exercise 10. *Estimate the total number of persons* $(\hat{Y})$ *in the population.*

Exercise 11. *Compute the standard error for* $\hat{Y}$.

Exercise 12. *Determine the coefficient of variation for* $\hat{Y}$.

Exercise 13. *Construct a 95 percent confidence interval for* $\hat{Y}$ *and interpret the result.*

CHAPTER 11

RATIO ESTIMATES

11.1 REASONS FOR CONSIDERING USE OF RATIO ESTIMATES

In earlier chapters we dealt with the problem of how to design the most efficient sample (from the point of view of minimizing the standard error) using as much relevant information as we can obtain about the population. We have seen how to use information for stratification with either proportionate sampling or optimum allocation, how to take unit costs into account, and how to choose between different kinds of sampling units. We have seen how to use whatever knowledge we have of costs and of the variances of different methods of sampling, in order to produce the maximum amount of information with the resources we have available. All of this analysis has been in terms of fairly simple estimates such as $\bar{y},\ \hat{Y},\ p\ \ (or\ \ \hat{P})$ in which the estimates were prepared by using only the sample data, the total number of units (N) in the population and the probabilities of selection. Thus, for simple random sampling,

$$\bar{y} = \frac{1}{n} \sum_i y_i$$

$$\hat{Y} = \frac{N}{n} \sum_i y_i = N\bar{y}$$

For stratified sampling

$$\bar{y}_{st} = \frac{1}{N} \sum_i \frac{N_i}{n_i} \sum_j y_{ij}$$

$$\hat{Y}_{st} = \sum_i \frac{N_i}{n_i} \sum_j y_{ij}$$

There are similar formulas for cluster sampling, or for estimation of proportions $(\hat{P})$.

There are, however, more complex methods of estimating these statistics, which under certain circumstances can result in very large reductions in the standard errors.

Moreover, there are other types of statistics which we wish to measure--such as ratios of two characteristics, change over time of a single characteristic, etc. For example, we may obtain information on wages and salary payments and on number of hours worked, but we may be more interested in estimating the average hourly earnings, rather than total wages and salaries or total hours worked. From surveys covering two different periods of time, we may be more interested in finding out whether total wages have gone up or down than in measuring the level at any one time. The analysis of the standard errors estimated ratios also helps with the problem of producing more efficient estimates of means and totals.

We shall investigate the simplest and most commonly used method of improving the reliability of an estimated mean or total, by the use of a special estimating technique which produces a "ratio estimate." A number of other very powerful tools are useful in particular situations; for example, difference estimates and regression estimates, double sampling (in which the final sample is selected from a previously selected larger sample that provides information for improving the final selection or the estimation procedure), and special methods for the estimation of time series. However, we will only discuss in this chapter ratio estimates.

11.2 RATIO ESTIMATES OF AGGREGATES

Ratio estimation is the most commonly used of the more complex estimation techniques available to the statistician. It is also the easiest to apply. It is appropriate whenever the units of the population possess two characteristics that are positively correlated--the higher the correlation, the greater the gain from using this technique. The simplest kind of ratio estimator of the form $\hat{Y}_R$ given by equation (11.1), is an estimate of Y (the population aggregate):[16]

$$(11.1) \qquad \hat{Y}_R = \frac{\hat{Y}}{\hat{X}} X$$

Here, $\hat{Y}$ and $\hat{X}$ are the ordinary estimates of the aggregates of two characteristics Y and X; the aggregate X must be known in order to estimate the aggregate Y.

To compute $\hat{Y}_R$ it is not necessary to compute $\hat{Y}$ $\hat{X}$ and since, for a self-weighting sample,

[1] The ratio estimator of a mean $\hat{\bar{Y}}_R$ (as an estimate of $\bar{Y}$) is obtained by dividing $\hat{Y}_R$ by N; it has the same coefficient of variation as the estimate.

$$\frac{\hat{Y}}{\hat{X}} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{1}^{n} y}{\sum_{1}^{n} x}$$

However, the formula in (11.1) is useful in deriving the variance of $\hat{Y}_R$.

Ratio estimates of aggregates are ordinarily applied in the three situations described in sections 11.2.1 to 11.2.3 below.

11.2.1 Ratio to Same or Related Characteristic at an Earlier Time Period

X is the same type of characteristic as Y, but X refers to an earlier time period during which a complete census was taken. For example, we may have taken a full census of manufacturers in one year, and wish to take a sample survey the following year. Suppose we wish to estimate the total value of shipments. For each manufacturing establishment in the sample, we obtain not only y_i, the value of shipments in the survey year, but also x_i, the value during the preceding census year. Then $\hat{Y}$ and $\hat{X}$ would be estimates from the sample of total shipments for the two years, obtained by the methods discussed earlier. X is the total value of shipments tabulated from the full census. In this application, the survey is actually used to measure the rate of change between the two years, using the identical sample of establishments. The rate of change is then multiplied by the census total for the previous year.

11.2.2 Ratio of Two Related Characteristics at the Same Time Period

Y and X are two different characteristics for the same time period, which are known to be positively correlated. The true value of the aggregate X is known. For example, for the i^th^ farm in a sample, x_i may be the total hectares in the farms, and y_i the payments for farm labor; the total hectares in all farms, X, is known from another source. If, in general, the larger farms pay more total wages for farm labor than the smaller ones, the ratio estimate can drastically reduce the sampling error. In this application, the survey is used to measure a rate (such as the average payment per hectare) which is multiplied by the known number of hectares.

11.2.3 Ratio of a Subset to the Total

The characteristic Y is a subset of X, varying roughly in proportion to X. For example, x_i may be total acres in the i^{th} farm in the sample, and y_i the acres planted to a particular crop on that farm. Another application is the case in which X is the total number of units of analysis and Y is the number of these having a particular attribute. For example, y_i might be the number of persons in the labor force in the i^{th} cluster; x_i is the total number[17] of persons in this cluster; and X is the known total number of persons in the population.

In these cases, the survey is used to measure a ratio $\hat{Y}/\hat{X}$ which is then multiplied by the population total (X) for the characteristic in the denominator of the ratio.

11.3 VARIANCE AND BIAS OF A RATIO ESTIMATE

In examining $\frac{\hat{Y}}{\hat{X}}X,$ it is clear that X is not derived from the sample. The sampling error in the estimate $\hat{Y}_R = \frac{\hat{Y}}{\hat{X}}X$ is, therefore, dependent on the sampling error of the ratio, $\hat{R} = \frac{\hat{Y}}{\hat{X}},$ with X having only the effect of a constant multiplier. Therefore, an analysis of the sampling error of $\hat{Y}_R$ is closely related to that of the ratio $\hat{R}$ $\frac{\hat{Y}}{\hat{X}},$ as an estimate of R = $\frac{Y}{X}.$

The mathematical form of the distribution of the ratio of two random variables from sample to sample is much more complicated than that of the simpler estimates discussed earlier. It involves the relationship of two variables, both of which have sampling errors. Hence, more care is required in deciding when to use such ratios. The following facts about the variance of ratios and ratio estimates will indicate when to use a ratio estimator

[2] *In cluster sampling, the estimate of the total number of units of analysis will be a random variable, which is usually not exactly equal to the true figure . Hence, the proportion of units having the attribute must be treated as a ratio of random variables.*

to estimate a mean or an aggregate. They also tell us what error to expect when using the estimate.

11.3.1 Variance of Ratios and Ratio Estimates

The variance of an estimated ratio $\hat{R} = \dfrac{\hat{Y}}{\hat{X}}$ is approximately

(11.2)
$$S^2(\hat{R}) = R^2 \left[\frac{S^2_{\hat{Y}}}{Y^2} + \frac{S^2_{\hat{X}}}{X^2} - 2\rho \frac{S_{\hat{Y}} S_{\hat{X}}}{YX} \right]$$

where R is the population ratio $\dfrac{Y}{X}$ (a ratio of aggregates), and

$$S^2_{\hat{Y}} = S^2(\hat{Y}) = N^2 \frac{(N-1)}{N} \frac{S^2}{n}$$

Similarly, the variance of the ratio estimate of a total , $\hat{Y}_R = \dfrac{\hat{Y}}{\hat{X}} X$ is

(11.3)
$$S^2(\hat{Y}_R) = X^2 S^2(\hat{R}) \qquad Y^2 \; [\frac{S^2_{\hat{Y}}}{Y^2} + \frac{S^2_{\hat{X}}}{X^2} - 2\rho = \frac{S_{\hat{Y}} S_{\hat{X}}}{YX}]$$

The alternative form of this equation is

(11.3a)
$$S^2(\hat{Y}_R) = \frac{N(N-n)}{n(N-1)} \sum (y_i - Rx_i)^2$$

and is estimated by ,

(11.3b)
$$s^2(\hat{Y}_R) = \frac{N(N-n)}{n(n-1)} (\sum y_i^2 - 2\hat{R} \sum y_i x_i + \hat{R}^2 \sum x_i^2)$$

Equations (11.2) and (11.3) are somewhat simpler if expressed in terms of the coefficient of variation, CV. The square of the coefficient of variation (that is, the rel-variance) of $\hat{R}$ is the same as that of $\hat{Y}_R$ and can be expressed as

(11.4) $$CV^2(\hat{Y}_R) = CV^2(\hat{R}) = CV^2(\hat{Y}) - 2\rho CV(\hat{Y})CV(\hat{X}) + CV^2(\hat{X})$$

In the above formulas, ρ is the coefficient of correlation between the variables Y and X. It represents the correlation of Y and X, not for the elementary units of analysis but for the units used for sampling. For example, if Y and X represent the incomes of persons in two different years, but the sample is a cluster sample, the correlation coefficient ρ will be the correlation between the values Y_i and X_i where Y_i is the sum of the incomes for all persons in the ith cluster in the year of estimation and X_i is the corresponding sum in the base year. Frequently, $\rho S_{\hat{Y}} S_{\hat{X}}$ is referred to as the sampling covariance between $\hat{Y}$ and $\hat{X}$ and the symbol $S_{\hat{Y}\hat{X}}$ is used for it. It can be calculated exactly as the variance, but with the cross product $(Y_i - \bar{Y})(X_i - \bar{X})$ replacing the square $(Y_i - \bar{Y})^2$ wherever it occurs. Thus, for simple random sampling we have

(11.5) $$\rho S_{\hat{Y}} S_{\hat{X}} = S_{\hat{Y}\hat{X}} = N^2 \frac{(N-n)}{N} \frac{S_{YX}}{n}$$

where

(11.6) $$S_{YX} = \frac{1}{N-1} \sum_1^N (Y_i - \bar{Y})(X_i - \bar{X})$$

and an estimate of S_{YX} can be made from the sample by using

(11.7) $$s_{yx} = \frac{1}{n-1} \sum_1^n (y_i - \bar{y})(x_i - \bar{x})$$

The corresponding estimate of ρ, designated by ρ’, is obtained by putting sample values of $S_{\hat{Y}}$, $S_{\hat{X}}$, S_{yx} and in place of the population values, in equation (11.5), and solving for ρ, which then becomes ρ' . The ρ' may also be computed directly from

(11.8) $$\rho' = \frac{\sum_{1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum_{1}^{n}(y_i-\bar{y})^2\ \sum_{1}^{n}(x_i-\bar{x})^2}}$$

For a stratified sample, with estimates of totals given by $\hat{Y}$ and $\hat{X}$

(11.9) $$S_{\hat{Y}\hat{X}} = \sum_{h=1}^{L} N_h^2 \frac{(N_h-n_h)}{N_h}\frac{S_{hYX}}{n_h}$$

where S_{hYX} are within-strata covariances and are computed in exactly the same way, but are restricted to the values within each stratum.

11.3.2 Gains with a Ratio Estimate

If we examine equation (11.4), the formula for the rel-variance of an estimate of a total,

(11.4) $$CV^2(\hat{Y}_R) = CV^2(\hat{R}) = CV^2(\hat{Y}) - 2\rho CV(\hat{Y})CV(\hat{X}) + CV^2(\hat{X})$$

we see that CV² of the ratio estimate $CV^2(\hat{Y}_R)$ can be expressed as CV² of the simpler estimate $CV^2(\hat{Y})$ plus the term $CV^2(\hat{X})$ minus the term $2\rho\, CV(\hat{Y})CV(\hat{X})$.
gain or lose by the use of a ratio estimate, as compared with the simpler estimate $(\hat{Y})$ depends on whether $(CV^2(\hat{X}) - 2\rho\, CV(\hat{Y})\,CV(\hat{X}))$ is smalle than zero. Another way of expressing this is the following:

(1) If $\rho > \frac{1}{2}(CV(\hat{X})/CV(\hat{Y}))$, a ratio estimate is more efficient

(2) If $\rho < \frac{1}{2}(CV(\hat{X})/CV(\hat{Y}))$, a ratio estimate is less efficient

(3) If $\rho = \frac{1}{2}(CV(\hat{X})/CV(\hat{Y}))$, both estimates have the same standard error.

11.3.2.1 High Correlation.

To see the implication of these facts in some common situations, consider the example of a census of manufacturers which was conducted in one year, followed by a sample the next year. Let y_i and x_i represent the values of shipments for the same sample firm in two consecutive years. In this case $CV(\hat{X})$ and $CV(\hat{Y})$ are nearly the same, and $(CV(\hat{X})/CV(\hat{Y}))$ is approximately 1.

Furthermore, there will be a very high correlation between Y and X, probably about 0.90 or 0.95. Consequently, a ratio estimate will result in a substantial gain in accuracy. The amount of the gain can be found as follows: if $CV(\hat{Y}) = CV(\hat{X})$, Equation (10.4) becomes

$$CV^2(\hat{Y}_R) = 2(1 - \rho)CV^2(\hat{Y})$$

and if $\rho = .90$, we have

$$CV^2(\hat{Y}_R) = 0.20\,CV^2(\hat{Y})$$

In other words, the use of a ratio estimate achieves an 80 percent reduction in variance. If $\rho = .95$, $CV^2(\hat{Y}_R)$ becomes equal to $[(0.10)*CV^2(\hat{Y})]$ and the reduction is 9 Looking at the result in another way, the ratio estimate is as effective as using a sample 5 times (or 10 times) as large.

11.3.2.2 Low Correlation.

Consider now the situation described in section 11.2.3 in which Y is a subset of X. In such cases, the correlation is likely to be quite low, unless $\frac{Y}{X}$ is fairly large--for example, greater than ½. In practice, if $\frac{Y}{X}$ is less than about 20 percent, a ratio estimate may increase the sampling error although, generally, not much. If $\frac{Y}{X}$ is greater than 40 or 50 percent, a ratio estimate will usually improve the efficiency; the closer to 100 percent, the more the improvement. Between 20 and 40 percent, the differences between the two types of estimates will be small. Thus, for example, in a labor force survey, the use of ratio estimates probably provides an important improvement in the estimate of the number of employed (which comprises a fairly high proportion of the adult population) but probably results in a slight increase in the standard error of the estimate of unemployed.

11.3.3 Bias of the Ratio Estimate

The ratio estimate is a biased estimate. This can easily be demonstrated by constructing a small population with values Y_i and X_i for each element, taking all possible samples of two or three elements, and computing $\hat{Y}/\hat{X}$ for each sample. It will be seen that the average of the ratios is not the true average. However, the bias tends to be negligible for moderately large samples. In most practical applications, the bias is so small compared with the advantage gained in reducing the sampling error, that the ratio estimate is preferred over the unbiased estimate.

11.3.4 Consistent Estimates

A ratio estimate, although biased, is a consistent estimate. This means that, if we use a large enough sample, we can be sure that the estimate will be as close as we like to the true value. Not only does the standard error decrease with increasing sample size, but the bias is also reduced.

11.3.5 Confidence Limits

For reasonably large samples, ratio estimates are normally distributed (for the kinds of populations dealt with in practice). Consequently, if we can compute the standard error of the ratio estimate, we can construct the same type of confidence limits

for $\hat{\bar{Y}}_R$ and $\hat{Y}_R$ $\bar{y}$ as for $\hat{Y}$; and that is, we can say that we have a 68-percent chance that a range around the estimate of plus and minus one standard error will cover the true figure, a 95-percent chance that a range of plus and minus two standard errors will cover the true figure, etc.

11.3.6 Minimum Sample Size Required

Sections 11.3.3 to 11.3.5 above refer to the fact that moderately large samples are needed to make the bias negligible, and to provide a reasonably normal distribution of sample estimates. When is the sample large enough? The following working rule has been suggested: If the sample size exceeds 30 and if the coefficients of variation of $\bar{y}$ and are both less than 10 percent, then the bias is negligible and we can assume that the theory for the normal distribution applies. The first condition does not mean that a ratio estimate is necessarily better than a simple unbiased estimate whenever n > 30; it means this size of sample is required before the formulas for sampling error have the usual meaning in terms of confidence intervals.

11.3.7 Formula for Bias

An approximation to the bias of an estimate of a ratio of two variables $\hat{R} = \frac{\hat{Y}}{\hat{X}}$ is

$$Bias \doteq R(CV^2(\hat{X}) - \rho CV(\hat{Y})CV(\hat{X}))$$

where ρ and R are defined as in section 3.1. For the estimate of a total, $\hat{Y}_R = \frac{\hat{Y}}{\hat{X}}X,$ the bias is

$$Bias \doteq Y(CV^2(\hat{X}) - \rho CV(\hat{Y})CV(\hat{X}))$$

Even with low values of ρ this will be small compared with the standard error of $\hat{Y},$ provided only that the sample is reasonably large so that $CV^2(\hat{X})$ is small.

These bias formulas are presented for analytical purposes. They are never used to adjust estimates. In situations where the bias would be expected to be significantly large, we would either increase the sample size or use a different method of estimation.[18]

11.3.8 Danger in Use of Ratio Estimate

If ratio estimates are applied separately for a large number of subgroups of the population, with a small sample in each subgroup, the bias in the subgroup may accumulate and become too large to ignore. For example, suppose a relatively small sample of persons is classified by separate age-sex groups--300 persons divided into 5-year age groups by sex. There would be about 30 such groups. Suppose we know the

[18] *For other estimation methods, see section 2 of chapter 11 of Sample Survey Methods and Theory (referred to in footnote 1 of chapter 8).*

true total population in each of these 30 groups. For any statistic we are interested in, we could compute a separate ratio estimate for the persons in each of the 30 groups, and then get a final estimate by adding the 30 results. The average size of sample in each group would be 10. Since there would be only a small sample in each of the age groups for which a ratio estimate would be formed, the accumulation of 30 different ratio estimates could result in a serious bias. In such a case, the use of ratio estimation group-by-group is not recommended.

11.3.9 Illustration

Suppose that a complete census of the value of manufacturing shipments was taken in 1981. The following table shows the value of shipments in each of a simple random sample of the value of 10 shipments drawn from the value of 30 shipments. The problem is to estimate the total value of shipments in 1982. The true 1981 total, X is assumed to be known . Its value is $19.5 billion.

Value of shipments in 1981 (x_i)	0.3	1.1	0.5	0.4	1.0	0.7	0.2	0.3	2.4	0.1
Value of shipments in 1982 (y_i)	0.1	0.6	0.8	0.6	1.0	0.8	0.9	0.8	2.7	0.2

We have,

N = 30, n = 10

$$\sum_{i=1}^{10} y_i = 8.5, \quad \sum_{i=1}^{10} x_i = 7.0$$

$$\sum y_i^2 = 11.79, \quad \sum x_i^2 = 9.1$$

Compute the estimate of the total and the variance, the coefficient of variation of the estimate and the confidence interval for Y by using (a) a method of simple random sampling and (b) a method of ratio estimates.

(a) Simple random sampling

(1) $$\hat{Y} = \frac{N}{n}\sum_{i=1}^{10} y_i = \frac{30}{10}(8.5) = \$25.5 \ billion$$

(2) $$s^2(\hat{Y}) = N^2\frac{(N-n)}{N}\frac{s^2}{n}$$

$$= 30^2\frac{(30-10)}{30}\frac{(.071)^2}{10} = \$30.43 \ \textit{billion}^2$$

(3) $$s(\hat{y}) = \sqrt{30.43} = \$5.52 \ \textit{billion}$$

(4) $$cv(\hat{y}) = \frac{5.52}{25.5} = 0.216$$

(5) 95 % confidence interval for Y is

$$25.5 \pm 1.96(5.52) = (\$12.42, \ \$38.58) \ \textit{billion}$$

(b) Ratio Estimates

(1) $$\hat{Y}_R = X\left(\frac{\sum_i y_i}{\sum_i x_i}\right)$$

$$= \hat{Y}_R = 19.5(\frac{8.5}{7.0}) = 19.5(1.21) = \$23.68 \ \textit{billion}$$

Using equation (11.3b),

(2) $$S^2(\hat{Y}_R) = \frac{N(N-n)}{n(n-1)} (\sum y_i^2 - 2\hat{R} \sum y_i x_i + \hat{R}^2 \sum x_i^2)$$

$$= \frac{30(30-10)}{10(10-1)} (11.79 - 2(1.21)(9.57) + 1.21^2(9.1^2))$$

$$= \$13.01 \ \textit{billion}^2$$

(3) $s(\hat{Y}_R) = \$3.61 \;\; billion$

(4) $cv(\hat{Y}_R) = 15.2\%$

(5) A 95 % confidence interval for Y is

($16.46, $30.90) billion

STUDY ASSIGNMENT

Problem *A: A 10-percent simple random sample of housing units in a village has been selected producing the 12 housing units listed below. At each sample unit, information was obtained on the number of persons in the household and the total annual earnings; the results are given below. It is also known from independent sources that the total population of all households in the village is 600 persons.*

Sample unit	*Total persons*	*Total earnings*
1	*6*	*$ 7,000*
2	*6*	*8,000*
3	*5*	*3,000*
4	*8*	*10,000*
5	*4*	*2,000*
6	*2*	*1,000*
7	*4*	*2,000*
8	*5*	*3,000*
9	*1*	*1,000*
10	*7*	*8,000*
11	*4*	*1,000*
12	*5*	*6,000*
Total	*57*	*$52,000*

Exercise 1. *Estimate the total earnings in all households in the village using a direct inflation factor.*

Exercise 2. *Estimate the total earnings in all households in the village using a ratio estimate.*

Exercise 3. *Use the sample results to estimate the coefficient of variation for each of the above estimates.*

Problem B: *The following table shows the total hectares in three farms along with the payments for farm labor draw1n from 30 farms. The true value of the total hectares of all farms, X is assumed to be 800.*

Farm *(i)*	*Hectares* *(x_i)*	*Payments* *(y_i)*
1	*5*	*382*
2	*8*	*467*
3	*10*	*701*

Exercise 4. *Estimate the total payments* $\hat{Y}_R$ *for all farms Y.*

Exercise 5. *Estimate the variance of* $\hat{Y}_R$.

Exercise 6: *Compute the coefficient of variation of* $\hat{Y}_R$.

Exercise 7: *Find a 95% confidence interval for Y.*

CHAPTER 12

SAMPLING FOR OBJECTIVE MEASUREMENT SURVEYS IN AGRICULTURE

12.1 NEED FOR OBJECTIVE MEASUREMENTS

The principles of sampling discussed in the previous lectures are widely applicable to survey programs generally. Certain kinds of surveys, however, may require special techniques of sampling and data collection which are determined by the nature of the inquiry or the ability of respondents to give accurate answers. Chapter 12 describes some special techniques used in agriculture surveys.

Statistics on area planted with individual crops and on yields from these crops are, in most countries, based upon periodic reports from crop reporters. In some countries, these reporters are holders or other individuals who reside in the rural areas and have knowledge of the local agriculture; they report voluntarily, usually by mail. In other countries the reporters are government officials or agents. The reports submitted by these agents are usually less accurate than those submitted by private individuals, in part because the agents are usually reporting for a much larger area and in part because the agents are not so closely connected with agriculture. However, whether made by private individuals or by government agents, these reports are all subject to biases which are often large and always difficult to evaluate. For example, investigations in various countries have shown that in estimating yields, reporters (particularly official reporters) have a tendency to be biased toward the normal; in other words, in good years they tend to underestimate the yield whereas in bad years they tend to overestimate. Although private reporters also have this tendency to some extent, they are generally more inclined to underestimate in the belief that it will be to their advantage to do so. Areas, on the other hand, tend to be overestimated because of the difficulty of making proper allowances for non planted areas around the edges of fields and areas within the fields that cannot be planted.

Check data from past years can be used to evaluate the biases in the estimates of production obtained from reporters. For crops such as tobacco or cotton, which must be processed before being used, information on production can be obtained from the processors and compared with the corresponding figures obtained from reporters. For other crops, similar use can be made of data obtained from marketing or shipping sources. If such data are complete (usually there is no guarantee that they are complete) and if the relative bias remains reasonably constant from year to year, estimates for the current year can be adjusted on the basis of this past experience. For other crops, which are at least partly consumed locally, fed to livestock, etc., such check data are not available. Census data, if available, can be used as a benchmark for adjusting the reports for these crops. However, the census data are also subject to reporting biases.

Furthermore, adjustments using census data become less and less reliable as the time lapse between the last census and the current year widens.
Experience in many different countries under a variety of conditions has indicated that subjective methods of estimating production, even when other data are available for adjusting the estimates, cannot provide reliable results. If accurate and unbiased estimates are required, the only alternative is to establish some type of program utilizing objective methods of observation applied on a random sampling basis. Such surveys are called "objective measurement surveys" because the data are collected by actual observation and measurement or counting, rather than by methods depending on the judgment, good memory, or education of persons who report the required information. Even though such a program of objective measurement surveys is relatively costly and difficult to carry out, the results will usually justify the effort.

12.2 DESIGNING THE SAMPLE

The theoretical considerations affecting sample design, discussed in previous lectures, are as relevant to the design of an objective measurement survey as they are to any other survey.

12.2.1 Type of estimates required

The sampling statistician must know whether estimates are required for the nation as a whole, for the Provinces or districts individually, or for some other administrative areas. The sample allocation must be planned to give estimates for the desired areas at an acceptable level of reliability. If an estimate of the number of holdings (either in total or for a specific crop) is also required, this must be considered in designing the sample.

12.2.2 Stratification

First-level strata often consist of the smallest areas requiring separate estimates. Further gains in efficiency may be obtained by further stratification into geographic areas having relatively homogeneous yield rates for the crop. Other bases for stratification, such as irrigated and nonirrigated land, varieties of crops, etc., may also be used.

12.2.3 Allocation to strata

The statistician must decide how to allocate the sample to strata. A common practice is to allocate it proportionately to the area under the particular crop or group of crops being investigated. If available, knowledge about the relative variances and/or the relative costs of performing the field work in the different strata should also be used in allocating the sample.

12.2.4 Sampling within strata

A decision must be made on the method of sampling within strata. As was indicated before, there are usually several possible sampling units and sample designs. In deciding upon a sampling plan, the sampling statistician will need to know what materials are available for constructing the sampling frame and what types of data are required. His choice may also be influenced by other factors such as the availability of capable personnel to carry out the work. However, even with the restrictions imposed by these considerations, there will usually be a number of possible choices.

12.2.4.1 Sample stages and types of sampling units

In most practical applications, several sampling stages and sampling units will be used within strata. For example, if the strata are large administrative divisions, such as Provinces, a sample of districts might be selected at the first stage and a sample of subdistricts within sample districts at the second stage. Where "villages" have identifiable boundaries and account for all the land, they can serve as convenient units at some stage in the sampling. The ultimate unit of analysis will usually be an individual holding, the individual field, or (for studies involving estimation of yields) small plots within fields. If the field is the unit of analysis, holdings may be selected at the preceding stage.

12.2.4.2 Methods of selecting holdings and fields

The following examples illustrate some procedures that can be used to select holdings and fields in the final stages of the sample design. The selection of plots within fields is discussed in section 12.4.4 of this chapter.

(1) Holdings can be selected from lists if lists are available or can be constructed without much difficulty. Lists of holdings would be needed only for the units (villages, subdistricts, etc.) actually selected in the sample at the preceding stage; if necessary, these could be compiled as part of the field operation. The selection of holdings can be made either with equal probability or with probability proportionate to size (assuming that information on size is available or can be obtained). The measure of size might be total reported area in the holding, total area in a particular crop or group of crops, etc.

Similarly, within each selected holding, a list of fields could be compiled and a sample selected. Again, selection could be made either with equal probability or with probability proportionate to size.

(2) If maps or aerial photographs are available, these can be used to select fields directly without first selecting holdings. One way to do this is to superimpose on the map or photo a grid on which dots have been placed either in a systematic pattern or at random; each field into which a dot falls is then included in the sample, thus giving the fields probabilities of selection proportionate to their sizes. This procedure requires, of course, that the maps or photos be sufficiently detailed so that the point and the corresponding field

can be located on the ground. (This procedure is not easily adaptable to estimating number of holdings, if that is desired.)

(3) Area segments are useful sampling units for determining which holdings and/or fields are to be included in the sample. These segments may be constructed either with natural boundaries that can be located on the ground or with imaginary boundaries drawn on a photo or map; the choice depends upon the particular situation. Holdings and/or fields may be associated with area segments in any of the following ways:

(a) Area segments with imaginary boundaries could be used as first-stage sampling units and a sample of segments selected; within the sample segments, fields could be selected as second-stage units in the manner described above in (2).

(b) An alternative procedure would be to include in the sample all fields (or holdings) for which a uniquely defined point falls within the segment boundaries. With this procedure, fields (or holdings) would not be selected with probability proportionate to their sizes; the probability of selection would be the same as the probability of selection of the segment into which the point falls. This is known as an open segment approach. The segments determine which units are included in the sample, but data are tabulated for some fields (or holdings) lying partly outside the segment and are not tabulated for other fields (or holdings) lying partly inside the segment.

The unique point must be defined with care. Usually a particular corner of the field (holding) would be designated as the unique point. Because fields (holdings) may not be rectangular, a specific rule for locating this corner would be needed as well. For example, if the northwest corner were the designated unique point, it could be defined either (1) by identifying the boundary points that lie farthest west and then designating the most northern of these points as the northwest corner or (2) by identifying the boundary points that lie farthest north and then designating the most western of these points as the northwest corner. If the holding were the unit of analysis, the residence of the holder (provided all such residences had a chance of being included in the sample) would generally be preferred as the unique point since it would be the easiest point to locate. A combination of rules is, perhaps, even more useful. For example, the residence of the holder might be used when the holder lives on the holding, and a particular corner used when he does not live on the holding. In any case, the point must be defined in a way such that it is truly unique (that is, each unit must have one, and only one, such point associated with it and thus have one, and only one, chance of being included in the sample); it should also be fairly easy to identify.

(c) If the unit of analysis is the holding, the weighted segment approach will usually be more efficient than the open segment approach. With this procedure, all holdings having any land in the segment are included in the sample. In the estimation, the data from each holding are weighted by a factor based on the proportion of the entire holding lying inside the segments. In almost all applications, the weighted segment approach requires that the segments have natural boundaries that can be identified on the ground.

(d) Still another possibility is to use the so-called closed-segment approach in which only those fields or parts of fields lying within the segments are included in the sample. One advantage of this procedure is that it avoids the difficulty of having to define the holding. Of course, if information is desired on a holding basis, the closed-segment approach is not appropriate since some holdings will certainly extend beyond the segment boundaries.

12.3 OBJECTIVE MEASUREMENT PROCEDURES FOR THE ESTIMATION OF AREA

Since it is known that data on land area obtained by asking individuals to respond to questionnaires can be very inaccurate, other means of obtaining these data have been investigated.[19] The usual approach in objective measurement surveys is to select a sample of areas, and then to go to these areas and measure them directly. There are also methods of obtaining objective estimates of area that do not require direct measurement of the land; for example, measuring the area on aerial photographs. In addition to the measurements, other information may be obtained. For example, the land may be classified into various categories according to its use (crop land, pasture, wasteland, etc.), the particular crop being grown on each piece of land may be identified, etc.

12.3.1 Measurement of land area

The first step in making direct measurements of land is to make a scale drawing. In order to do this, one must be able to measure distances and angles. A drawing made by a professional land surveyor using technical equipment would be very precise. On the other hand, a drawing made by an inexperienced worker measuring distances by pacing and measuring angles by eye estimates would not be very accurate. Between these extremes, there are many other methods that can be used. One should balance the relative cost against the relative accuracy of the various procedures and select the method that will provide an acceptable level of reliability for the lowest cost.

After the scale drawing has been made, the area of the drawing must be determined. If the land that was measured is in the shape of a regular geometric figure such as a

[1] For discussion of techniques and experiences in many countries, see S. S. Zarkovich (ed.), Estimation of Areas in Agricultural Statistics, Food and Agriculture Organization of the United Nations, Rome, 1965.

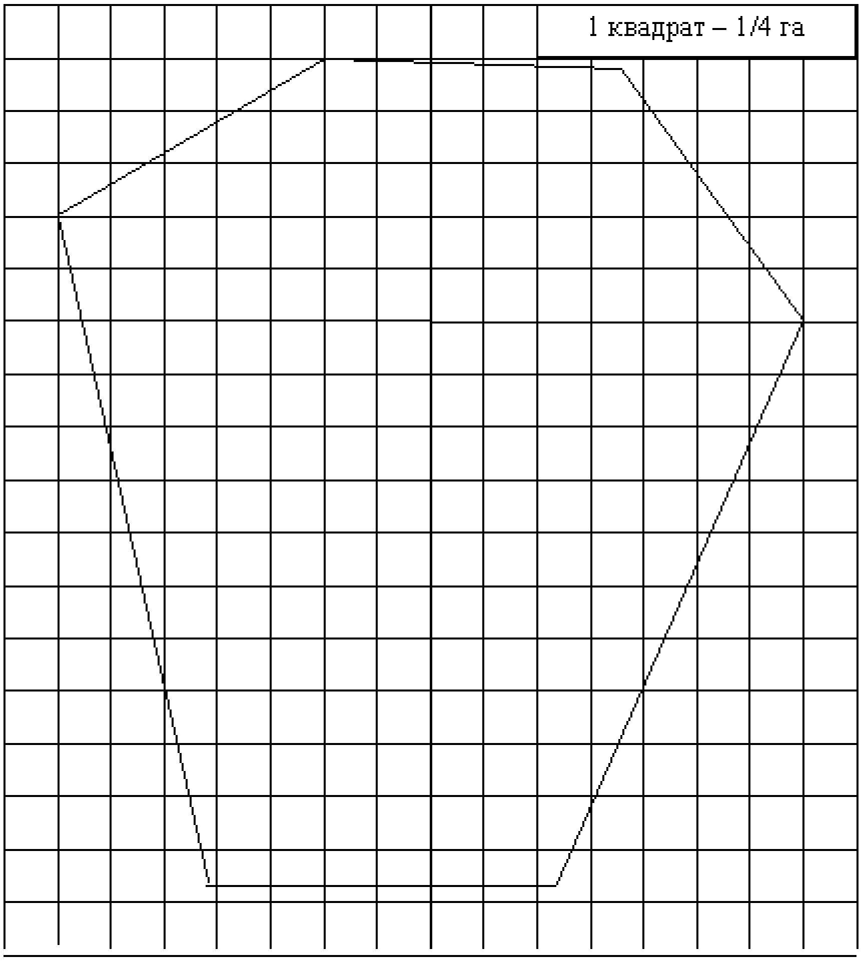
rectangle, trapezoid, etc., it is relatively easy to determine the area of the drawing by standard mathematical formulas. Using the appropriate expansion factor, the area of the land represented by the drawing can then be determined. Often, however, the area is of irregular shape and other methods must be used; for example, triangulation, planimetering, gridding, dot counting, and map cutting and weighing.

12.3.1.1 **Triangulation.**--In triangulation, the polygon formed by the drawing is converted into simple triangles. It is a principle of geometry that this can always be done. (Curved boundaries are roughly approximated by a series of straight lines before triangulation.) Each triangle is measured and the area computed by standard formulas. This procedure is time consuming and tedious and has largely been replaced.

12.3.1.2 **Planimetering.**--A planimeter is an instrument with which one can determine the area of a closed figure by tracing around the boundary of the figure with a pencil-like device. A good planimeter will give very accurate results. It does, however, require a skilled operator and much time.

12.3.1.3 **Gridding.**--Basically, a grid is a plane divided into small squares (for example, a piece of ordinary graph paper). For use in measuring area, the squares are constructed so that each is equivalent to a particular amount of area in accordance with the scale of the drawing. A transparent plastic grid can be placed over the drawing; or the grid can be printed on paper and the drawing made directly on this paper. To estimate the area represented by the drawing, one counts the whole squares and parts of squares within the perimeter of the scale drawing and converts this number to its equivalent in terms of the appropriate unit of area.

Figure 1: MEASUREMENT BY GRIDDING
(1 SQUARE = 1/4 HECTARE)


1 квадрат – 1/4 га

Although not as accurate as planimetering, gridding can be done in less time. It requires only that the individual be able to count accurately and that he be able to accurately convert the partial squares into an equivalent number of whole squares. See Figure 1 on the preceding page for an illustration of this method. There are approximately 159 squares within the scale drawing (including the partial squares that overlap the boundary); thus, since each square represents 1/4 hectare, the field contains about 40 hectares.

12.3.1.4 **Dot counting**. Dot counting is essentially the same as gridding except that instead of small squares, the grid consists of uniformly spaced dots. Each dot represents a unit area according to the scale of the drawing. One need only count the dots lying within the perimeter of the drawing to find the area. If any dots lie on the boundary, only half of them are counted.

12.3.1.5 **Map cutting and weighing.**--By this procedure, the map or photograph of the area is carefully cut into pieces representing different categories of land along the lines drawn by the field worker. Each piece is then carefully weighed. The estimation is based on the weight of the paper in each category relative to the weight for the entire area. This procedure is not very practical; it is time consuming and requires a weighing instrument of high precision and map paper of uniform quality.

12.3.2 Observation of land uses for a sample of points or lines

Some methods of objectively measuring area do not require direct measurement of the land itself. Instead, the proportion of land falling into various categories is estimated by some objective means and multiplied by the known total area of land in the universe (Province, district, etc.) to estimate the total area in each category. All of the methods discussed in section 3.2 except the last method (the last method described in paragraph 3.22) require accurate, up-to-date maps or aerial photographs; consequently, their usefulness is somewhat limited at this time. However, as progress is made in aerial photography, these and similar methods are likely to become more generally useful in the future.

12.3.2.1 **Observations for a sample of points.**--A sample of points is selected and the points marked on maps or aerial photographs. In selecting the sample of points, appropriate techniques of stratification and clustering should be used to maximize the efficiency of the design. For example, if primary interest is in the estimation of crop areas, higher sampling rates should be used in those portions of the universe known to consist primarily of crop land.

If only broad categories of land use are to be estimated, and suitable aerial photographs are available, it may be possible to make the necessary observations directly from the photographs. For most purposes, however, it will be necessary to send observers to the field to locate each sample point and to record the crop being grown or other use being made of the land at the point.

One author has suggested that for periodic surveys the sample points be permanently identified by suitable markers, to make them easier to locate. The markers could not be placed at the exact locations of the sample points, since they would interfere with farming operations; however, they would be placed nearby and equipped with sighting devices aimed at the sample points. This method has not yet been tried in the field. (Refer to "Fixed-Point Sampling--A New Method of Estimating Crop Areas" by Thomas B. Jabine in Estadistica, published by the Inter-American Statistical Institute, Washington, D.C., September-December 1967.)

Once the observations have been made for the sample of points, one can make an unbiased estimate of area devoted to a particular use:

(1) For each stratum in which points were sampled at a constant rate, tally the number of sample points in each land use category.

(2) Multiply the known total area of the stratum by the proportion of sample points devoted to that use.

(3) Sum over all strata.

12.3.2.2 Observations for a sample of lines.--A sample of lines is selected and the lines are marked on maps or aerial photographs. As in the case of points, appropriate techniques of stratification and clustering should be used to increase the efficiency of the design. The usual procedure within ultimate sampling units is to select a sample of parallel lines spaced at equal intervals.

By using aerial photographs, or by actually pacing the lines, the investigator determines the proportion of each line falling into each land use category. Unbiased estimates are then made from these observations by a procedure completely analogous to that described above for point samples.

A relatively cheap but biased form of line sampling involves the substitution of roads for a probability sample of lines. The investigator drives a car along a prescribed route. The car is equipped with a distance measuring device. As he drives, the investigator notes and records the distance for which the road is bordered by each category of land being measured (specific crops, crop land in general, pasture, woodland, etc.). Estimates are then made in the normal way for line sampling.

This last technique is likely to be seriously biased, especially in areas where the road network is sparse, since the pattern of land use along roads is likely to differ substantially from the overall pattern for a given area. Techniques based on probability sampling should be used in preference if at all possible.

12.3.3 Use of ratio estimation and double sampling to improve efficiency

Having completed area measurements on the holdings (or other units of analysis) in the sample, we can estimate totals directly from these data by the estimation procedure which is appropriate to the particular sample design. This procedure can usually be improved upon, however, if in addition to making area measurements for a sample of the population, we also have available less accurate and less expensive area data (for example, data obtained by direct interview) from the entire population. Such data would normally come from a complete census. By means of ratio estimation, we can often obtain estimates of population totals that will be more reliable than those that could be obtained from either the objective measurements or the interview responses alone. The procedure is essentially the same as that discussed in section 2.3 of chapter 10. The X-characteristic in this case would be the actual measurement of the land obtained for a subset of the population; the Y-characteristic would be the data collected by the interview.

Even more useful and practical is a technique called double sampling[20] in which the less expensive technique is used to obtain data from a relatively large sample of the population and the more expensive technique to obtain data from a subsample of the basic sample. Again, ratio estimation is used, but here the Y-characteristic is the response that is obtained by the less expensive technique, and the sample estimate of the population total for the Y-characteristic is used in place of a total based on 100-percent coverage.

Compared with the method based on area measurement alone, methods using ratio estimation will be preferred if the gain in efficiency more than offsets the cost of obtaining the supplementary observations by the less expensive technique (either from the entire population or, in the case of double sampling, from a larger sample from the population). The factors to be considered are:

(1) The strength of the relationship between the data obtained by the two methods. The interview response must have a high positive correlation with the area measurement if a significant improvement is to be obtained. One would reasonably expect this to be the case.

(2) The relative cost of the two methods. Assuming that the correlation is large enough, ratio estimation will reduce the number of holdings requiring area measurement in order to achieve a given level of reliability. Whether or not this reduction will offset the cost of obtaining the interview responses depends in part upon the difference in costs between the two types of observations.

Compared with the method based only on interview responses, the use of ratio estimation will be preferred whenever it is believed that the bias in the interview responses is sufficient to justify the additional expense of obtaining the area measurements. The

[2.] Double sampling is a statistical technique useful in a variety of situations whenever a characteristic of interest that is difficult or expensive to determine is correlated highly with another characteristic that can be determined relatively easily or inexpensively.

concept of mean square error (MSE) is needed to understand the situation more fully. Recall from previous chapters that the variance is based on differences between estimates (x') based on samples and the value X that would be obtained if data had been collected from all members of the population, using the same techniques. The mean square error, on the other hand, is based on differences between estimates based on samples and the true value of the quantity being measured (X_T). If the data-collection technique is unbiased, $X = X_T$, then the MSE is equivalent to the variance; if the technique is biased, the MSE is equal to the variance plus the square of the bias ($X - X_T$), or

$$(12.1) \qquad \text{MSE} = S^2(\hat{X}) + (X - X_T)^2 \qquad .$$

For a given cost, data can be obtained by interview from a sample of a certain size. For the same cost, data can be obtained by interview from a smaller sample, combined with objective measurements from a subsample of this sample. Estimates based on the large interview sample will have a specified MSE containing a bias component as well as a variance component. Ratio estimates based on the combination of interview and objective measurement data will have a smaller bias but a larger variance. The MSE may be either larger or smaller than the MSE based only on the large interview sample depending on the variability in the population, the relative cost of the two procedures (which determines the relative sample sizes), the relative size of the biases (or the effectiveness of the ratio estimation procedure in reducing the bias), etc. The sampling statistician must consider all of these factors in allocating the available resources between the two procedures. His goal is to minimize the MSE for a given cost (or to minimize the cost of obtaining an acceptable level of reliability).

12.4 OBJECTIVE MEASUREMENT OF YIELD

The goal of objective measurement of yield is usually to estimate the yield of a crop on a unit basis (such as bushels per acre, quintals per hectare, etc.). In order to estimate the total production, it is necessary to have also an estimate of the total area of the crop in question planted. In some instances, only the yield is estimated by objective means, although estimates of both the yield and the area should be based on objective measurements.

The general procedure in making objective measurements of yield (usually called "crop cutting") is to use a random process to select areas (usually called plots) planted, and to cut and weigh the produce from each of these plots at or near the time the remainder of the field is harvested.[21] Each different crop has different characteristics, and the same crop will behave differently in different parts of the world. Consequently, there is no specific set of rules that can be applied to all crops or even to the same crops in different locations. We will, however, discuss in general terms some of the factors to be considered in planning such a program and describe some of the techniques that have

[3] Objective measurements are also used to forecast yields on the basis of observations made earlier in the season. Since the sampling procedures used in forecasting yields are quite similar to those used in estimating yields, only the latter are discussed in this section.

been used in the past.

12.4.1 Pilot studies

Because information gained about other crops or about the behavior of the crop in question in other countries is not directly transferable to one's own situation, pilot studies should be carried out before establishing any program for objective measurement of yield. Pilot studies can provide important information about most of the things that need to be considered such as sampling variability, optimum size and shape of plot, harvesting procedures, problems such as personnel and materials needed to carry out the work, etc. They are also useful as training devices for those who will eventually be in charge of the full-scale operation. On the basis of the pilot studies, the investigator can develop a sampling plan and field procedures appropriate to the conditions under which the survey will be conducted.

After a procedure has been decided upon, it is usually advisable to put it into operation only gradually and, after it is in full operation to carry it out for a few years simultaneously with the procedure it is to replace. The existing program, no matter how inadequate it may be, should not be ended until the proposed new method has been sufficiently tested and found to be clearly superior and operationally feasible.[22] After its superiority and feasibility have been established, the new method can then serve as a basis for evaluating the bias in the old method which would not be possible unless the two were conducted simultaneously for a few years. This is particularly important to users of the data who are interested in examining differences or trends over a period of years; they must know to what extent observed differences in the data are simply the result of differences in measurement technique.

12.4.2 Variability

One must have some idea of the variability in yield of the crop to be measured in order to plan wisely. Two aspects of variability which are of interest are:

(1) The relative variability of yields for different sizes and shapes of plots.

(2) For a plot of given size and shape, the relative magnitude of the variation among fields and the variation among plots within a field.

In deciding which type of plot to use, the investigator must balance the variability against the cost. He will attempt to select the plot that will give the desired degree of reliability for the lowest cost, although other factors (for example, personnel considerations) may force him to choose one that is quite the best in terms of costs and variances.

Experience has shown that in almost all cases, the variation among fields is considerably

[4] Actually, it may be necessary to continue the existing program in any case, particularly if data are required for administrative areas different from those for which estimates are made using objective data. Furthermore, the existing program may collect data on a number of crops which are not economically important enough to justify an expensive objective measurement program.

greater than variation within fields. As a result, the number of plots selected within each sample field should be small so that the available resources can be more efficiently expended on sampling as many different fields as possible. In fact, in some investigations, the optimum number of plots has been only one per field.[23] A minimum of two plots is necessary, of course, if one wishes to estimate the within-field variability from the sample; nevertheless, the investigator may choose to have only one plot per field if the within-field component of variance is very small compared with the between-field component.

12.4.3 Size and shape of plot

Circular, triangular, square, and rectangular plots have all been used in past studies for crops that are scattered in the field or planted in very closely spaced rows (for example, small grains or hay). For crops in widely spaced rows (for example, maize or cotton), rectangular plots are the logical choice; the width is often designated in terms of rows and the length in terms of feet (or meters, etc.).

Along with the shape of the plot, a method of marking it must be specified. Rigid frames or other devices have been used successfully for marking small plots. Ropes, chains, etc., are easier to transport but are more difficult to place in the field if the worker has to measure and drive stakes at the corners, etc. For a triangular plot, a closed chain with rings at the three vertices can be used quite easily; the same device, provided it forms a right triangle, can also be used to mark rectangular plots using a suitable combination of triangles. Large plots are usually laid out using pegs or stakes, string, and a measuring tape.

As the size of the plot increases, the variability among plots decreases; however, since the within-field contribution to the overall variance is usually negligible relative to the other sources of variance, small plots are usually preferred from a practical standpoint. One man can usually do the work alone, he can place a portable frame much faster than he can stake out a large plot, he can harvest more quickly, and he has less material to handle.

Unfortunately, experience has shown that small plots almost always produce seriously biased estimates. The reasons for this are not entirely clear, but it appears that two factors are largely responsible:

(1) In locating the plot in the field, it is much easier for the field worker to allow the condition of the crop to influence the precise location of the smaller plot.

(2) The problem of whether to count plants on the boundary as being in or out of the plot is more critical with the smaller plot, since the perimeter of a small plot is greater relative to its area than is the perimeter of a large plot. The general tendency appears to be to include plants that should be excluded and, thus, to consistently overestimate the yield. For a smaller plot, even a single plant erroneously included can seriously affect the results.

[5] Theoretically, the optimum number of plots need not be an integer. As a practical matter, of course, the theoretical result must be rounded to an integer.

12.4.4 Locating the plot in the field

Many different procedures have been proposed for locating plots in the field. Whatever method is used, it is important that the field staff understand clearly how it should be done, and checks should be made to see that they are following the instructions. Otherwise, subjective bias on the part of the field worker will almost certainly enter into the procedure.

Ideally it would be desirable to divide the entire field into plots of the size and shape decided upon and select the required number of plots at random. However, this is not usually practicable. A method that has been used and is practicable whenever the field is rectangular (or can be conveniently enclosed in a rectangle) is to locate points at random within the field; the sample plots are then laid out in a prescribed manner about these points. For each plot to be located, the procedure is as follows:

(1) The field worker selects a random number x between 0 and n_1, where n_1 represents the total length of one dimension of the field (or of the enclosing rectangle); he selects another random number y between 0 and n_2, where n_2 represents the total length of the other dimension. For a row crop, the first dimension would usually be expressed in terms of the number of rows.[24] In other cases, the dimensions would be expressed in terms of units, such as meters, or in terms of steps or paces.

(2) Starting at a predetermined corner, the field worker measures or paces (or counts rows) the distance x along the appropriate side of the field (or of the enclosing rectangle); then at right angles to this side, he measures or paces the distance y into the field.

(3) If the worker is still within the boundaries of the field, he marks the random point (for example, by digging with his heel and driving a stake). If he is not within the boundaries of the field (he would, of course, be within the enclosing rectangle), he uses another pair of random numbers and repeats the process.

(4) From this point, the field worker lays out the plot. If the plot is to be circular, the random point should be used as the center. If it is to be triangular or rectangular, the point should be used to locate a predetermined vertex or corner; this vertex or corner is usually chosen so that the plot will extend away from the random point in the direction that the worker has been walking.

Figure 2 on the following page illustrates this procedure. In this example the point (x_1, y_1) falls inside the field and is accepted. The point (x_2, y_2) falls outside the field and is rejected. From the sample point, the plot would usually extend upward and to the right.

[6] The random number would then be selected between 1 and the total number of rows in the field (n_1)

Figure 2: LOCATION OF RANDOM POINTS WITHIN A FIELD

One difficulty in this scheme is that it allows plots to overlap field boundaries; any of the several feasible rules that can be used in such cases present certain problems. Consider, for example, a field of maize 200 rows wide and 100 meters long. Suppose that the plot is to be 4 rows wide by 6 meters long. Suppose further that the selected row coordinate is 198 and the length coordinate is 95. From the point of intersection of the coordinates, the plot would extend 1 meter and 1 row beyond the boundaries of the field (the plot starts at the end of te 95th meter but includes row 198). Possible rules that could be adopted to take care of this situation include:

(1) Instruct the worker to harvest only the partial plot 3 rows by 5 meters and, of course, to record these dimensions on his form. Using the proper inflation factor, an unbiased estimate of the yield for this field could be made. In this example, this procedure could be carried out rather easily; however, if the field were irregular in shape or the plot were circular or triangular, the worker might find it difficult to estimate the portion of the plot in the field.

(2) Instruct the worker to think of the rows as being numbered in a circular manner

and similarly the length. Thus, in this example, row 1 would be the fourth row of the plot and the first meter in each row would be taken to finish out the length of the plot. This, too, would be an unbiased procedure. It would, however, not be practicable for anything except rectangular plots in regularly shaped fields. Furthermore, it might be difficult to explain it to the average field worker. Finally it does not fit into the usual concept of a plot as a contiguous piece of land.

(3) Instruct the worker to restrict his random selection to numbers that will not allow this situation or, equivalently, to reject plots found to overlap boundaries and select another set of coordinates. In this case, in the example, he could do the former by restricting the selection for rows to numbers between 1 and 197 and for length to numbers between 0 and 94. This procedure is clearly biased since the edges of the field (in the example, the first and last four rows and the first and last six meters) have less chance of being in the sample than does the remainder of the field. If the yield tends to be greater or smaller than average around the edges of the field, estimates of yield based on this method will be biased. However, this is the simplest procedure. If the borders of the field are small in area relative to the remainder of the field or if there is no reason to believe that the yield is different along the edges, this method can be recommended in preference to unbiased but more difficult procedures.

12.4.5 Harvesting procedure

If the plots are small, the field worker will probably do the work himself, cutting the crop and weighing it in the field. He will than take a small subsample to be sent to the central office for drying. (It is always a good practice to return the remainder of the produce to the holder.) If plots are large enough, it may be desirable to harvest them by the same method that the holder will use in the regular harvest and, if possible, at the same time. This will require his cooperation and help.

12.4.6 Adjustment to actual production

The technician's method of harvesting small plots and processing the produce usually gives a higher rate of yield than does the normal harvesting procedures used by the holder because of greater harvesting losses in the normal methods. For some crops, these losses are substantial. In addition, it is not possible to harvest all plots on or immediately before the harvest date. If the worker waits too long to start harvesting, he will almost certainly find some fields harvested before he arrives; consequently, he will need to start harvesting plots in some fields while the crop is immature. Both of these factors will cause biased estimates if adjustments are not made.
(The harvesting of small plots measures what is often referred to as biological yield.)

One method of adjustment is to select a subsample of fields of known area and harvest them for the holders, using the normal procedures. This provides a basis for adjusting the data collected from the harvested plots. A similar method appropriate for some crops (for example, hay crops that are taken from the field in the form of bales) is to arrange to weigh te entire crop in a subsample of fields as the holder transports it from the harvested field, but allowing the holder to harvest it whenever and however he wishes.

Another method of adjustment is to carry out a gleaning operation after harvest to estimate field losses directly. The estimated field losses per unit area are then subtracted from the estimated biological yield to get the actual yield. This procedure has the advantage of not requiring the worker to be present at the harvest--an important consideration since several holders of different sample fields may all decide to harvest on the same day. Unfortunately, experience has shown that the problems of estimating field losses are fully as great as those of estimating the original biological production.

As already mentioned, it is desirable that sample plots be harvested as near as possible to the date the remainder of the field is harvested; however, this cannot always be accomplished for all fields. One object of a pilot study would be to determine what adjustments, if any, must be made for differences between these harvesting dates. For many crops, no adjustment is necessary because the crop has essentially completed its growth before either date and is then in the process only of losing moisture.

An additional adjustment that must be made is for moisture content. A procedure commonly used is to dry the material from the plots (or a subsample of it) until it is at or very near to 0% moisture content and then to weigh it. This so-called dry weight can then be adjusted to any moisture content desired. For many crops, a standard moisture content has been specified. If the dry material is only a subsample of the plot, a two-step process is required. The material from the entire plot and the subsample must be weighed separately in the field immediately after cutting. The subsample is then dried and weighed. The dry weight of the entire plot can then be estimated using the ratio of dry to wet weight of the subsample.

12.4.7 Operational considerations

Before an extensive program to measure yields objectively can be put into operation, numerous practical problems must be solved. These include the availability of labor, the availability of facilities for drying the crops, equipment needs, the need to coordinate the activities of the workers with the holders' plans for harvesting their crops, etc. The problem of timing can be very difficult, particularly when the crop is likely to be ready for harvest at the same time over a wide area. As stated previously, one important reason for conducting pilot studies is to obtain information about these practical problems.

Study Assignment

Problem A. *The sketch below simulates a segment outlined on an aerial photo.*

The segment contains a total of 100 hectares divided into categories according to the uses made of the land. The categories are:

Crop land:

A1 - maize
A2 - wheat
A3 - other crop land

B - grassland
C - forest
D - wasteland

A grid of 36 dots has been placed over the segment to be used in estimating the amount of land by categories of use.

Exercise 1. *Estimate the number of hectares in this segment that are used for crop land.*

Exercise 2. *Estimate the number of hectares for grassland.*

Exercise 3. *Estimate the number of hectares in forest and wasteland.*

Exercise 4. *Estimate the proportion of crop land used for maize. In what basic way does this estimate differ from those in exercises 1 to 3?*

Problem B. *In the sketch above, marks on the east and west boundaries of the segment subdivide the boundaries into 40 units. Using these marks as guides, place two lines at random across the segment parallel to the north and south boundaries.*

Exercise 5. *Use these parallel lines to estimate the quantities estimated in Problem A.*

Exercise 6. *For each quantity, compile the distribution of the estimates obtained by several trials or several persons.*

Problem C. *The sketch below shows a field bordering on a river.*

Exercise 7. *Draw a circle around the corner corresponding to the unique point according to each of the definitions given below. Place the appropriate letter (a, b, c) by each circle.*

(a) Northwest corner - Identify those boundary points lying farthest north. The northwest corner is the most western of these points.

(b) Northwest corner - Identify those boundary points lying farthest west. The northwest corner is the most northern of these points.

© *Southwest corner - Identify those boundary points lying farthest south. The southwest corner is the most western of these points.*

Problem D. *Data on the total area of crop land harvested has been obtained by interview from a simple random sample (selected without replacement of 24 holdings out of a population of 96 holdings. Objective measurements have been carried out on a subsample of 8 of these holdings selected at random without replacement. The data are shown in the table below.*

Unit	Hectares of crop land harvested	
	Interview (Y)	Objective measurement (X)
1	14	14.4
2	79	-
3	46	-
4	112	116.1
5	46	-
6	92	-
7	29	-
8	40	41.9
9	12	-
10	78	80.4
11	66	-
12	43	-
13	39	-
14	91	93.9
15	17	16.8
16	68	-
17	100	-
18	87	-
19	74	75.4
20	64	-
21	78	-
22	40	42.6
23	22	-
24	55	-

Exercise 8. *Estimate the total crop land harvested using the interview data only. Estimate the variance of this estimated total.*

Exercise 9. *Estimate the total crop land harvested using the objective measurement data only. Estimate* *the variance of this estimate.*

Exercise 10. *Using the formulas given below, estimate the total crop land harvested and the variance of this estimate using both types of data and ratio estimation.*

$$Y = N\left(\frac{\bar{y}_2}{\bar{x}_2}\right)\bar{x}_1$$

$$s^2(\hat{Y}) = N^2\left[\left(1 - \frac{n_2}{n_1}\right)\frac{s^2(\hat{R})}{n_2} + \left(1 - \frac{n_1}{N}\right)\frac{s^2(y)}{n_1}\right]$$

where n_1 = size of large interview sample

n_2 = size of objective measurement subsample

$$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i$$

$$\bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$$

$$s^2(\hat{R}) = s^2(y) - 2\hat{R}s(yx) + (\hat{R})^2 s^2(x)$$

$$s^2(y) = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y}_2)^2$$

$$s^2(x) = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_i - \bar{x}_2)^2$$

$$s^2(yx) = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y}_2)(x_i - \bar{x}_2)$$

$$\hat{R} = \frac{\bar{y}_2}{\bar{x}_2}$$

SELECTED LIST OF REFERENCES

1. Cochran, William G. ***Sampling Techniques.*** Second edition. New York, John Wiley and Sons. 1963.

2. Food and Agriculture Organization of the United Nations (FAO). ***Estimation of Areas in Agricultural Statistics.*** Edited by S. S. Zarkovich. Rome, 1965.

3. Food and Agriculture Organization of the United Nations (FAO). ***Estimation of Crop Yields.*** By V. G. Panse. Rome, 1954.

4. Food and Agriculture Organization of the United Nations (FAO). By S. S. Zarkovich. ***Sampling Methods and Censuses.*** Rome, 1965. ***Quality of Statistical Data.*** Rome, 1966.

5. Hansen, Morris H.; Hurwitz, William N.; and Madow, William G. ***Sample Survey Methods and Theory.*** New York, John Wiley and Sons, 1953. (Volume I: Methods and Applications; Volume II: Theory)

6. Kish, Leslie. ***Survey Sampling.*** New York, John Wiley and Sons, 1965.

7. Kniceley, Maurice R. ***Probability Sampling for Surveys and Censuses,*** Course Notes, PSDP, 1985.

8. Megill, David J. ***Preliminary Recommendations for Designing the Master Frame for the Senegal Intercensal Household Survey Program,*** U.S. Bureau of the Census, November 1990.

9. Neter, John and Wasserman, William. ***Fundamental Statistics for Business and Economics.*** Boston, Mass., U.S.A., Allyn and Bacon, 1961.

10. Sampford, M. R. ***An Introduction to Sampling Theory.*** Edinburgh and London, Oliver and Boyd, 1962.

11. Sukhatme, Pandurang V. ***Sampling Theory of Surveys with Applications.*** Ames, Iowa. U.S.A., The Iowa State College Press, 1953. New Delhi, India, The Indian Society of Agricultural Statistics, 1953.

12. The RAND Corporation. ***A Million Random Digits.*** Glencoe, Illinois, U.S.A., The Free Press, 1955.

13. United Nations.Statistical Office. ***Handbook of Household Surveys: A Practical Guide for Inquiries on Levels of Living.*** New York, 1964. (Studies in Methods, Series F, No. 10)

14. U.S. Bureau of the Census. ***The Current Population Survey Reinterview***

Program, Some Notes and Discussion. Washington, D.C., U.S. Government Printing Office, 1963. (Technical Paper No. 6)

15. U.S. Bureau of the Census. ***The Current Population Survey--A Report on Methodology.*** Washington, D.C., U.S. Government Printing Office, 1963. (Technical Paper No. 7)

16. U.S. Department of Commerce. ***Statistical Abstract,*** Washington, D.C., U.S. Government Printing Office, 1981, Table 202, P. 123.

17 Yates, Frank. ***Sampling Methods for Censuses and Surveys.*** Third Edition. New York, Hafner Publishing Company, 1960.