United States General Accounting Office

GAO

Program Evaluation and Methodology Division

Revised
May 1992

Using Statistical Sampling

GAO/PEMD-10.1.6

Preface

GAO assists congressional decisionmakers in their deliberative process by furnishing analytical information on issues and opinions under consideration. Many diverse methodologies are needed to develop sound and timely answers to the questions that are posed by the Congress. To provide GAO evaluators with basic information about the commonly used methodologies, GAO's policy guidance includes documents such as methodology transfer papers and technical guidelines.

The purpose of this methodology transfer paper on statistical sampling is to provide its readers with a background on sampling concepts and methods that will enable them to identify jobs that can benefit from statistical sampling, to know when to seek assistance from a statistical sampling specialist, and to work with the specialist to design and execute a sampling plan. This paper describes sample design, selection and estimation procedures, and the concepts of confidence and sampling precision. Two additional topics, treated more briefly, include special applications of sampling to auditing and evaluation and some relationships between sampling and data collection problems. Last, but not least, the strengths and limitations of statistical sampling are summarized. The original paper was authored by Harry Conley and Lou Fink in April 1986. This reissued version supersedes the earlier edition.

Using Statistical Sampling is one of a series of papers issued by the Program Evaluation and Methodology Division (PEMD). The purpose of the series is to provide GAO evaluators with guides to various aspects of audit and evaluation methodology, to illustrate applications, and to indicate where more detailed information is available.

We look forward to receiving comments from the readers of this paper. They should be addressed to Eleanor Chelimsky at 202-275-1854.

Werner Grosshans
Assistant Comptroller General
Office of Policy

Eleanor Chelimsky
Assistant Comptroller General
for Program Evaluation and
Methodology

Contents

Figures

Abbreviations

FPC	Finite population correction
GAO	U.S. General Accounting Office
ICC	Interstate Commerce Commission
LTPD	Lot tolerance percent defective
PEMD	Program Evaluation and Methodology Division
PPS	Probability proportional to size

Chapter 1

Introduction

Sampling is a very important element in the design and planning of a job here at GAO as well as any evaluation. The purpose of this paper is to help GAO managers and evaluators learn more about statistical (or probability) sampling and the role it plays in the design and execution of a job. We have attempted to take the mystery out of what is often thought of as an esoteric subject by "walking" the reader through the various sampling procedures. In this document, we have chosen to describe the computations and sample selection procedures as they are normally done by computer in GAO, but formulas necessary to do the necessary calculations by hand can be found in any sampling textbook.

This paper makes the assumption that the reader has had a one-semester college course in statistics. However, those who have not had such a course (or who think they may have forgotten the basics) can refer to the second chapter and to appendix I. We do not expect that after reading this paper, the GAO evaluator will be able to design and carry out a statistical sampling plan without assistance. Rather, we hope to provide enough background on sampling concepts and methods to enable evaluators to (1) identify jobs that can benefit from statistical sampling, (2) know when to seek assistance from a statistical sampling specialist, and (3) work with the specialist to design and execute a sampling plan.

As in evaluation design, sample designs are characterized by the manner in which the evaluators have defined and posed the evaluation questions for the study, developed a statistical approach for answering those questions, formulated a data collection plan that anticipates problems, and detailed an analysis plan for answering the study questions with appropriate data.

Sampling is nothing new or unusual. For thousands of years, people have been basing judgments about a large group of objects on their observations of a few of them. Prehistoric humans probably decided whether the berries on a bush were edible by tasting a few of them (with possibly fatal results). At harvest time, farmers judged the quality and expected yield of a wheat field by rubbing the husks off a few ears of grain pulled from various parts of the field. People have used sampling techniques such as spot checking for many years. The great improvement in the last hundred years or so has been the development of statistical sampling. We now have ways of drawing and analyzing samples to produce more objective information of better quality and of being explicit about its limitations. Sampling is one aspect of GAO assignments and, consequently, the design of a sample is one part of an overall assignment design. The time to start consideration of sampling is during job design.

Evaluation Design

The design of any job starts with the question being asked. In Designing Evaluations (transfer paper 10.1.4 issued in May 1991), audit or evaluation questions are described as descriptive, normative, and impact (or cause and effect).[1] (See the "Papers in This Series" section.) The answers to descriptive questions provide information on existing conditions. The answers to normative questions compare (for ease of explanation) an observed outcome (this type is not limited to outcomes) with an intended level of performance. The answers to impact questions indicate whether observed conditions, events, or outcomes can be attributed to program operations. The methods used to answer evaluation questions, known as audit or evaluation strategies, can also be classified. As discussed in Designing Evaluations, the

[1]Here we refer to the few broad questions that dictate an evaluation's objective; later we will be concerned with the much narrower issues that must be addressed in designing a sample.

strategies and the types of questions most commonly addressed by each strategy are shown in table 1.1.

Table 1.1: Four Strategies and Their Evaluation Questions

Audit or evaluation strategy	Type of question most commonly addressed
Sample survey	Descriptive and normative
Case study	Descriptive and normative
Field experiment	Impact (cause and effect)
Use of available data	Descriptive, normative, and impact (cause and effect)

In a sample survey, data are collected from a sample of a population (some textbooks and statisticians use the word "universe") to determine the prevalence, distribution, or interrelationship of events and conditions. The case study analytically describes an event, a process, an institution, or a program; this strategy can use either a single case or multiple cases (see transfer paper entitled Case Study Evaluations, listed in "Papers in This Series"). The field experiment compares outcomes of program operations with estimates of what the outcomes would have been in the absence of the program. The use of available data refers to the use of previous reviews or data bases previously collected and still relevant.

No matter which strategy is used, evaluators need to consider several elements in designing a job. Designing Evaluations lists seven design elements:

1. kind of information to be acquired,

2. sources of information (for example, types of respondents),

3. methods to be used for sampling sources (for example, random sampling),

4. methods of collecting information (for example, structured interviews and self-administered questionnaires),

5. timing and frequency of information collection,

6. basis for comparing outcomes with and without a program (for impact or cause-and-effect questions), and

7. analysis plan.

In this paper, we are concerned primarily with the methods used for sampling information sources. Although we briefly discuss data collection methods, two other transfer papers in this series, Developing and Using Questionnaires and Using Structured Interviewing Techniques, describe these two methods in much greater detail.

Sample Design As an Element of an Assignment Design

In the context of auditing and evaluation, a sample is a portion of a population of possible information sources, and sampling refers to the methods for selecting those sources. Sampling is an element of the assignment design and, along with such other elements as data collection and analysis methods, determines the soundness of the answers to our evaluation questions.

The broad sampling options available may be understood by an example. Suppose we want to know how federal centers for runaway youths are operated. For certain kinds of information (for example, project costs and staff size), the directors of runaway-youth centers are probably the best source of information. We may then regard the center directors as our population of possible information sources. An important job design issue is how to select the directors from whom we will seek the information we need.

Three options are available for choosing the directors. One possibility is to gather information from all center directors. This is called a census, and it may be thought of as a special case of sample—a sample of all possible information sources. Sometimes, conducting a census is a desirable course, but it is not the main subject of this paper.

A second possibility is to apply a judgmental process to the selection of center directors. We might look at the locations of the centers and choose directors so that cities, suburbs, and rural areas are each represented to some degree. Or, bearing in mind the cost of travel to the center sites, we might choose the directors who are located closest to our office. Judgment sampling, which can be used to select information sources in many different ways, is largely outside the dimensions of this paper.

The third possibility is to select center directors by statistical (or probability) sampling. Here, chance determines which directors are selected. Most of this paper is devoted to statistical sampling and to describing the variety of ways in which chance can be involved (sample design), the processes for choosing information sources (selection procedures), and the methods for drawing conclusions about the population based on information about a sample (estimation procedures).

Representativeness: The Goal of Statistical Sampling

In many GAO jobs, the objective is to answer questions about a population of people or things. In the example of centers for runaway youths, we wanted to know how the centers were operated. This objective can be achieved by looking at a sample of centers, if the sample is representative of the population.

A representative sample has approximately the same distribution of the characteristics of which we are measuring as the population from which it was drawn. A detailed discussion of the concept of a representative sample is outside the scope of this paper but most people have an intuitive understanding of representativeness (see Kruskal and Mosteller, 1979 and 1980, for an extensive treatment). A representative sample of runaway-youth centers is like the population in terms of characteristics such as number of center staff, types of youths who come to the center, average duration of stay, and so on. With such a sample, we infer that the characteristics of the population, which we do not know, are like the characteristics of the sample, which we do know.

Statistical sampling produces a sample that, it can be persuasively argued, is representative of a population. However, samples of a population differ from one another as well as from the population itself. Hence, it is desirable to have an objective measure of the possible variation between samples and of the sample's relationship to the population. With this information, it is possible to determine the amount of error that arises because our sample does not correspond exactly to the population. This is an important feature of statistical sampling. It allows us to be precise about the amount of error introduced by the sampling process. We can then decide whether the amount of error is tolerable when weighed against trade-off factors, such as the cost of obtaining a larger sample that will have less error.

Random Selection

The essence of statistical sampling is selecting a sample by some random (or chance) process. By randomizing the sample selection, we make sure that the sample represents the population within the limits of some measure of the imprecision induced by sampling, and we can measure the precision of the information yielded by the sample.

The term random selection does not mean a haphazard or "catch as catch can" sample, such as inspecting poison gas shells that are stored closest to the entrance of an ammunition bunker or interviewing "average-looking" people in shopping centers. Rather, to select randomly is to eliminate personal bias or subjective considerations from the selection of the sample items. Every item in the population has a known probability of being selected, and the selection of an item does not affect the selection of any other item. If one were to draw different samples from the same population, the results would differ for each sample, but these differences would stem from chance, not personal bias or other systematic factors.

The selection of a sample by some random method in order to obtain information or draw conclusions about the population of interest is referred to as probability or statistical or scientific sampling. Regardless of the name used to describe the method, the key elements are that (1) each element in the population has a known (nonzero) probability of being selected and (2) the actual selection technique truly executes the random method.

Distinguishing Between Samples and Examples and Between Sampling Operations

It is important to distinguish between samples and examples. A statistical sample, as we have stated, is selected in a way such that the information obtained represents the characteristics of the population from which it was selected. An example, however, assists in describing findings and recommendations or demonstrating a particular point. Usually, the characteristics of an example are known before it is selected. The example may be selected as a typical case, or it may be selected to represent an unusual or problem situation. Examples can be chosen from the items that were already selected in a probability sample. There is no objection to using carefully selected items as examples, provided that we describe them as examples and do not imply that they are representative of the population of interest.

It is also important to distinguish between three interdependent sampling operations: sample design, selection procedures, and estimation procedures. Each of these is discussed in detail in chapter 3. Sample design refers to the plans made for the overall way in which a sample will be related to a population.

The sample design affects the estimation procedures used, and it may also affect the selection procedures used. Conversely, the sample design is often affected by the estimation procedures the evaluators want to use. Further, selection procedures can have a major effect on how the measure of error (called precision) is estimated, and the types of estimates to be developed can have a bearing on the selection procedures to be used.

The Organization of This Paper

Three sampling strategies are available: census, judgment sampling, and statistical sampling. This paper focuses on statistical sampling. If statistical sampling is part of the job design, the further choice of a particular sampling procedure, such as simple random sampling or cluster sampling, is necessary. And to implement the assignment design, two other major steps are required: sample selection and estimation of the population's characteristics.

Chapter 2 provides a review of the basic concepts of statistics and the basic formulas that form the foundation of sampling. Chapter 3 provides an overview of all three components: sample design, sample selection, and estimation. It also provides more detail on statistical sample designs by covering simple random sampling, stratified sampling, and cluster sampling. Chapter 4 takes up the matter of basic estimation, using the concepts of confidence and precision. Chapter 5 contains more advanced estimation procedures. Some special sampling issues that apply more to auditing than to evaluation are discussed in chapter 6. Chapter 7 discusses sample selection—the considerations in randomly selecting the sample units. Chapter 8 provides a bridge between sampling and topics on data collection and analysis, such as missing data and nonresponses. Chapter 9 briefly summarizes the strengths and limitations of statistical sampling.

Some topics of a more technical nature appear in the appendixes, as follows. Appendix I discusses some of the theory that underlies sampling. Appendix II presents a comprehensive description of sampling procedures. Appendix III discusses the computations used for stratified and one-stage cluster estimation. Appendix IV lists various packaged, or "canned," computer programs that can do sampling computations.

A Review of Basic Concepts

A primary objective of statistics as used by the GAO evaluator is to describe a population. How many units are involved? What are the most common values? What range of values can we expect to encounter? Knowledge about these features can often provide clues about where to concentrate an evaluation effort. In many instances, data on the entire population may be available through the agency's computerized records. When such information is available, it should be presented so as to provide as accurate a description of the population as possible. Often a frequency distribution is the best way to do so, because statistics by their very nature present an incomplete and potentially misleading description of the population.

A frequency distribution is a representation of the number of times the members of a population fall into a category. This category must be exclusive—that is, each member of the population can belong to one and only one category. Many computerized statistical packages are available for determining data frequencies.

For example, in a recent GAO report, we classified fixed benefit pension plans according to their industry group, as given in table 2.1.

Table 2.1: A Frequency Distribution

Industry group	Number
Wholesale trade	2,114
Retail trade	1,579
Finance, insurance, and real estate	1,060
Legal, medical, and health services	6,821
Other services	3,270

Statistics can be thought of as the shorthand we use to describe the population. They provide a quantifiable way to describe the distributions of the information we have gathered. Because statistics represent a summary or shorthand description of the data, they tend to distort the "truth."

Statistics: A Shorthand Description of a Population

Statistics can be either calculated directly from the population data (statisticians call these parameters) or estimated from sample data. In general, two major statistics are used to describe a frequency distribution: (1) measures of central tendency, which tell us what we can expect a typical or middle data point to be, and (2) measures of dispersion. Dispersion refers to the spread of the data—that is, the extent to which the information is scattered.

Measurement Levels of Data

Before describing measures of central tendency and of dispersion, we should mention the concept of levels of measurement. The measurement level of a data point reflects the ordering or distance properties inherent in the measurement scale. The traditional measurement classification system identifies four levels: nominal, ordinal, interval, and ratio.

A knowledge of the levels of measurement and their implications is important to the users of statistics because each statistical technique is appropriate for the data measured only at certain levels. The computer does not know what level of measurement underlies the numbers it receives, and it will process whatever numbers are entered into it. Thus, it is up to evaluators to determine whether a particular technique is suitable for their data. A brief discussion of each of the measurement levels follows.

Nominal

This is the "lowest" of the measurement levels. As the word "nominal" implies, this measurement level involves simply sorting units into unique classifications, by assigning a name or label to each one. Our aim is to sort them into categories that are similar, often with the hope that they will be similar with respect to other data values as well. For example, we might categorize people according to gender. At this level of measurement, no assumption of ordering between the categories is made.

Ordinal

In some cases, the measurement categories may have been ordered according to the degree to which they possess a characteristic, even though we cannot say how much of it they possess. Ordinal measures manifest all the features of the nominal level, with the addition of an order. Military rank is one example. Another appears in questionnaires, where a common measurement tool is the five-point scale strongly agree, agree, neutral, disagree, and strongly disagree.

Interval

At the interval level, we know not only the order of categories but the magnitude of difference between them as well. For example, we know that the difference between 35 and 40 degrees on a thermometer is the same as the difference between 80 and 85 degrees. The important thing to note is that the interval scale does not have an inherently determined zero point (zero is determined by an agreed upon definition). By this we mean that while the difference in both cases is 5 degrees, we cannot say that 80 degrees is twice as hot or twice as cold as 40 degrees. Consequently, this scale allows us to study the differences between values but not their proportionate magnitudes.

Ratio

The ratio measurement level has all the properties of the interval scale plus a natural zero point. Common ratio scale measurements are weight, distance, speed, and monetary value. Hence, it is meaningful to say that if I have $200 and you have $100, then I have twice as much money as you do.

Central Tendency

One of the first things people generally want to know about a collection of data points is its "average" value. They want a "sense" of what to expect if the data were represented by a single number. This concept can be expressed statistically in several ways.

Arithmetic Mean

The mean is the total of all the values for the items divided by the number of items. For example, we may suppose there are 10 computers at an Air Force location, and we wish to summarize the number of days of downtime for these computers. The data are given in table 2.2.

Table 2.2: Days of Downtime for 10 Computers

Computer number	Days
1	7
2	23
3	4
4	8
5	2
6	12
7	6
8	13
9	9
10	4

Thus, to calculate the mean we would add the days of downtime and divide by the number of computers. In this case the mean would equal (7 + 23 + 4 + 8 + 2 + 12 + 6 + 13 + 9 + 4)/10, or 8.8 days during the year. The formula is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $\bar{x}$ = mean, Σ = the sum of, x_i = the individual values for the sample or population, and n = the sample size or population size.

The mean, as a single number representing a whole set of data, has important advantages. First, its concept is familiar to most people and intuitively clear. Second, every data set has a mean. We should also mention at this point that while every data set has a mean, not every one is meaningful—for example, mean of gender might be (1 = female, 2 = male) 1.5. The mean is a measure that can be calculated and that is unique because every data set has one and only one mean. Third, every observation in the data set is taken into account when we calculate the mean. As a result, the mean is a reliable measure, less likely to be determined by chance than by some other characteristics of the data set.

Like any statistical measure, however, the mean has disadvantages, as well as advantages. First, while the mean is reliable in that it reflects all the values in the data set, it can be affected by extreme values that are not representative of the rest of the data. For example, if we were to calculate the mean income of five people whose incomes were $100, $100, $100, $100, and $100,000, then the mean would be $20,800, which does not reflect that four of the five had incomes of $100. A second problem with the mean is that it is tedious to compute, because we use every data point in the calculation. Finally, we are unable to compute the mean for a data set that is categorized in such a way that one or more categories are open-ended.

Median

The median is the midway numerical value if the values are arranged in size order. Therefore, this measure of central tendency applies to ordinal or higher levels of measurement. It is a single value that measures the central item of the data set. This single item is the middlemost or most central value in the set of data. Half of the data set lies above this point, and the other half lies below it.

The median has several advantages over the mean. The most important is that extreme values do not affect the median as strongly as they do the mean. For example, the median income of the five people mentioned above would be $100. The median is easy to understand and can be calculated from any kind of distribution of the appropriate scale, even grouped data with an open-ended category, unless the median falls into that category.

The median has some disadvantages as well. Statistical procedures that use the median for inferring information about the population are more complex than those that use the mean.

Mode

The mode is the most frequently occurring numerical value, or the category that has the greatest number of the units in the population belonging to it. For example, in the frequency distribution in table 2.1, the category containing legal and medical services is the mode class, since the largest number of pension plans occurred in this category.

The mode, like the median, can be used as a measure of central tendency for qualitative as well as quantitative data. In fact, the mode can be used as a measure of central tendency for all four of the levels of measurement. Also like the median, the mode is not unduly affected by extreme values. Even if the high values are very high and the low values very low, we choose the most frequent value of the data set to be the modal value. We can use the mode no matter how large, how small, or how spread out the values in the data set happen to be. Another advantage of the mode is that it can be used even when one or more of the categories are open-ended.

Despite these advantages, the mode is not used as often to measure central tendency as are the mean and median. Too often, there is no modal value because the data set contains no values that are repeated. Other times, every value is the mode since every value occurs the same number of times. Clearly, the mode is a useless measure in these cases. Another disadvantage is that when data sets contain two, three, or many modes, they are difficult to interpret and compare.

Comparing the Mean, Median, and Mode

When we work with statistics, we must decide whether to use the mean, the median, or the mode as the measure of central tendency. Symmetrical distributions that contain only one mode always have the same value for the mean, median, and mode. In these cases, we need not choose the measure of central tendency because the choice has been made for us.

In a distribution in which the values are concentrated at the lower end of the measurement scale, the mode has the lowest value, the median is higher, and the mean is the highest value. The reverse is true if the values are concentrated at the high end of the scale. In these cases, the median is often the best measure of central tendency because it is always between the mean and the mode. The median is not as highly influenced by the frequency of occurrence of a single value as is the mode, nor is it pulled by extreme values as is the mean.

Otherwise, there are no universal guidelines for applying the mean, median, or mode as the measure of central tendency for different populations. Each application must be judged independently.

Dispersion

Another common question about a collection of data points is "how far do they spread out from the center?" Are the data all tightly grouped about the measure of central tendency, or do they vary a great deal? The usual measures of dispersion are described below.

Range

The distance (or difference) between the highest and lowest values is the range. For example, if the highest value for a given set of data were 27 and the smallest value were 1, then the range would be 27 - 1, or 26. This is a quick measure of the dispersion of the distribution.

Semiquartile Range

This is a measure similar to the procedure that is used in scoring some events in the Olympics. The first step in computing the semiquartile range is to order the data items from highest to lowest. We then find the quartiles—that is, we divide the data into quarters. Thus, the lowest quartile is the value marking the lowest 25 percent of the values; the highest quartile is the point marking the highest 25 percent of the data. This range is then defined as half the difference between the lowest and highest quartiles. For example, if the upper quartile were 20 and the lower quartile were 4, then the semiquartile range would be (20 – 4)/2, or 8. This method is not used very often, but it can be useful for making a quick estimate of the spread. It is also a quick method for measuring the dispersion of the data around the median.

Mean Absolute Deviation

The deviation is a measure of the difference between the individual items in a population and the mean value. The mean absolute deviation is simply the average of the total unsigned differences. (If the differences were signed, the total of these differences would be zero.) This measure is hardly ever used because if the deviations (or differences) are known, then it is easy to compute other measures of dispersion. The formula for computing the mean absolute deviation is

$$MAD = \frac{\Sigma\, [x - \overline{x}]}{n}$$

where MAD = mean absolute deviation, Σ = the sum of, [ ] = absolute value, x = individual value, $\overline{x}$ = mean, and n = the sample size.

Variance

This measure is sometimes called the average squared deviation. It is computed by taking the difference between individual value and the mean and squaring it. Then you add all the squared differences and divide by the number of items. The formula for computing variance is

$$V = \frac{\Sigma\, (x - \overline{x})^2}{n}$$

where V = variance, Σ = the sum of, x = individual value, $\overline{x}$ = mean, and n = the sample size.

Standard Deviation

The standard deviation is the square root of the variance. We take the square root to account for the fact that we squared the differences in computing the variance. This is probably the most common and useful of the dispersion measures.

For an example of the last three measures of dispersion, let us look at the data in table 2.3.

Table 2.3: Data for Measures of Dispersion[a]

Value	Mean	Difference	Absolute difference	Squared difference
3	5	−2	2	4
4	5	−1	1	1
5	5	0	0	0
6	5	1	1	1
7	5	2	2	4
25	**25**	**0**	**6**	**10**

[a]Mean = 25/5 = 5. Signed deviation = 0. Unsigned deviation = 6. Mean absolute deviation = 6/5 = 1.2. Variance = 10/5 = 2. Standard deviation = square root of 2 = 1.41.

Coefficient of Variation

The coefficient of variation is the ratio (expressed as a single number) produced by dividing the standard deviation by the mean value. The coefficient of variation for the data set in table 2.3 is 0.282, or 28.2 percent. It provides an indication of the consistency of the data. (See table 2.4.)

Table 2.4: Interpreting the Value of the Coefficient of Variation

If the coefficient of variation is	Data have
Less than 30%	Small variation
30%-49%	Moderate variation
50%-69%	Medium variation
70%-89%	High variation
90% or more	Extreme variation

Some Comments on Looking Summary Statistics in the Eye

While all the measures of central tendency and dispersion that we have talked about are useful indicators, the evaluator must be very careful in applying such measures to real data. Don't forget to look at the raw data. This may seem like a simple requirement, but, unfortunately, there are many cases where it is not possible. For example, when the data set is very large, looking at the raw data tables is not very useful.

Insist on seeing the actual data whenever you can. The number of faulty data points discovered in many listings is usually quite surprising to the data analyst. Errors in recording the data, errors in collecting the data, omission or duplication of the data, all these can occur. Sometimes they can be spotted by studying the raw data, sometimes not. Their existence is a consistent hazard since all the statistics in the world will not correct these kinds of mistakes.

As far as summary statistics are concerned, each statistic that can be used has its own peculiar applicability. The only way we can judge its usefulness in summarizing the important characteristics of any data set is to examine the data.

An Overview of Sample Design and Selection and Estimation Procedures

Sample Design

Sample design is a part of the overall assignment design composed of the seven elements listed in chapter 1. Designing a job is an iterative process involving these several elements. We must tentatively formulate the evaluation questions and then make preliminary decisions about sampling issues and data collection methods. It is also advisable to have a preliminary data analysis plan in mind before making the final sample design.

A sample design documents the steps and procedures involved in taking a sample. It guides evaluators in executing the sample and aids in preparing the scope and methodology section of the report. Sometimes the sample design is called a sampling plan. Sample design (or a sampling plan) involves the following steps:

- formulating the objectives for the assignment;
- stating the sample objectives (for example, to estimate the number of tax returns for which the government owed the taxpayer interest);
- explaining the reason for taking a sample rather than a census;
- defining the sampling unit, or the elements or objects on which the measurements will be made (for example, tax returns, heads of households, or participants in a program);
- defining the population of interest, including an estimate of its size (for example, the 150,000 tax returns handled by the ABC service center during March 1991);
- developing the sampling frame, or the physical list, or where not reasonable to obtain the physical list, then a description of the items available for selection in the sample (for example, the computer list on tape of all returns processed during the month; note that this may not be the same as the population of interest, since some items of interest may not be obtainable);
- describing the type of sampling to be done and the reasons why this method was selected (for example, a

stratified sample was selected from the personnel in each of the armed services because we wanted to make estimates not only across the Department of Defense but also for each service). If a judgmental sample is to be used, explain why;

- describing the sample selection procedure used in selecting the sampling units, including the source of the random numbers;
- stating the required confidence level (GAO normally uses a 95-percent confidence level);
- suggesting a sample size and the precision you expect to achieve (many times these may not be known until after a preliminary sample is taken and analyzed);
- deciding the data collection and recording techniques to be used to record the data;
- choosing the analysis methods to be used; and
- explaining how missing sample items and outliers will be handled.

While these are steps in the sampling plan, they do not fall into the nice neat pattern that is implied here. Each of these steps is necessary, but they may need to be taken through several iterations and consultations with both the subject matter and the technical specialists.

Defining the Population and the Sampling Units

It is necessary to define the population very carefully, because this is the entire collection or group of items to which our estimates and inferences apply. In many projects, more than one population of information sources will be of interest. In the example of the runaway-youth centers, it may be desirable to obtain information not only from center directors but also from staff members, youths staying at the centers, and the parents of the youths. In principle, the sampling considerations in choosing directors are simply extended to the other populations but, in practice, some designs may be more advantageous than others. Sampling specialists should be consulted.

The logical starting place may be with either the population or the sampling units. We might begin, for example, with the understanding that we want to draw conclusions about the population of rail shipments of ammunition in 1990. We must then decide upon the sampling units. Do we want to define the sampling unit as the total shipments made by a depot, the total shipments received by a military unit, the government bill of lading for an individual shipment, or something else?

If we begin with a sampling unit defined as a government bill of lading involving rail shipments of ammunition, then we must be clear about just what population we want to draw conclusions about. Do we want to include shipments from all ammunition depots in the country or Army depots only, and are the shipments for an entire year, a single month, or a quarter?

Once the population has been defined, we must either obtain or develop a sampling frame. The sampling frame is a list or method of obtaining the items in the population. The list can be printed on paper, it can be a magnetic tape file or a file on a computer disk, it can be on microfiche, or it can be a file of accounts-receivable ledger cards or stock record cards. The frame should have several characteristics. First of all, the frame should permit the sampler to identify and locate the specific item that is to be drawn into the sample and to differentiate this item from all other items in the sampling frame. The frame should also contain all the items in the population. For example, if the population has been defined as the civilian work force at a naval shipyard, then the list of workers from which the sample was drawn should have included all civilian workers on the date of the audit or evaluation, should contain no duplicate entries, and should not contain entries that are not in the population. However, it should be noted that the frame may not contain all the population; the difference between the population of interest and the sampling frame is called the sampling gap. For example, if we were going to conduct a survey of all adults in the country and we were to do the interviews by telephone, the sampling gap would be the adults in the country who were not available by telephone. It is not necessary to literally "list" the universe. For example, it is possible to randomly select from the list of all possible telephone numbers without possessing a physical list of such numbers. Sometimes the list exists only in a conceptual sense.

In addition, we may want to define subdivisions of the population. One type of subdivision is the stratum, a subpopulation obtained by dividing the population into two or more mutually exclusive groups, or strata, which we can do if we know in advance the number of sampling units in each stratum. Independent random samples are selected from each stratum in order to obtain more precise estimates or to emphasize certain portions of the population, such as units with a high dollar value or a great potential for error. Often, the stratification system is based on the locations of the sampling units in the population. Examples of strata are households classified as urban, suburban, or rural; naval bases classified by geographic location; and taxpayers classified by income.

Another type of population subdivision is the domain of interest. This type of subdivision is necessary when separate estimates are needed for each of a number of classes into which a population may be divided but we do not know in advance the number of sampling units in each group. Thus, we must depend on the sample if we are to develop this information. Examples of domains of interest are students at a university who intend to major in education, travel vouchers involving the use of personally owned vehicles, and farms worked by tenants.

The sampling units are often defined to be persons or things we want to study—the units of the population about which we need information. But sometimes, because of the arrangement of the population, the lack of a list of the items we want to observe, and practical considerations, we may have to select a sampling unit that is larger than the unit about which we want to obtain data. An example is selecting a household in order to determine the employment or health status of its members. In this example, the item of interest, the household member, is called the secondary sampling unit, and the larger unit, the household, is called the cluster or primary sampling unit.

The primary sampling units must (1) be mutually exclusive and (2) constitute the entire population or include the entire population of secondary sampling units. This means that each unit being observed, the secondary sampling unit, must belong to one and only one primary sampling unit and that the primary sampling units must contain or cover the entire population of interest.

Sometimes the cluster or primary sampling unit consists of so many items that we must select a sample of items within each primary sampling unit. This is called two-stage cluster sampling. Occasionally, it is necessary to select a sample of primary units, a sample of secondary units from within each primary unit, and a sample of items from within each secondary unit. This is called multistage cluster sampling.

Sometimes, samples are taken in two or more phases, or "waves." This technique may be used to take a large preliminary sample, classify the sample into two or more domains of interest, and then draw smaller subsamples from the domains of interest. This type of sampling is known as double or two-phase sampling. An excellent example of double sampling cited by

Cochran (1977) involved surveys of the German civilian population in 1945, when the sample from each town was usually drawn from rationing lists. It was proposed that the population be stratified by age and sex. Because the sample had to be drawn in a hurry and the rationing lists were in constant use, it was not possible to tabulate the population by age and sex. However, a moderately large sample of names could be selected quickly. Each person selected was classified into the appropriate sex-age class. From these classifications, smaller samples of persons to be interviewed were selected.

Another type of two-phase sampling is drawing repeated samples from the same population. The usual purpose of these samples is to measure change from a preceding time period or periods and to obtain current estimates on various statistical measures of interest. The general procedure is to replace part of the sample (or select new sample units) and retain part of the sample every time the data are collected. An example is the current population survey conducted jointly by the Bureau of the Census and Bureau of Labor Statistics to measure employment and unemployment. In this survey, one fourth of the households are replaced by new sample households each month, so that a household is in the survey for 4 months. The household is omitted from the survey for the following 8 months, brought back into the survey for 4 more months, and then dropped. The current population survey uses this procedure because (1) more accurate measures of change are obtained by looking at differences in the same units over time and (2) respondent burden is limited by restricting the number of periods any one household can be included. This type of sampling is also called panel survey.

Choosing a Sampling Strategy

Three broad sampling strategies were outlined in the first chapter. The choice of a census, a judgment sample, or a statistical sample is a job-design decision of great importance. Besides the evaluation objective, factors such as cost, precision, and the feasibility of drawing certain kinds of samples must be considered. Although this paper is primarily about statistical sampling, a brief outline of the pros and cons of the different sampling strategies is appropriate. Before a sampling strategy is designated, the evaluation staff should seek assistance from their appropriate design, methodology, and technical assistance group or from the sampling statistician in PEMD.

Census

For some GAO projects, a census is appropriate, as when the individual items in the population are very important in themselves or when the information to be obtained is critical and the population is small enough to enable the evaluators to survey all the units within their resources. On other occasions, the population may be so small that sampling is not needed. Also, when all the data are already on a computer or in some machine-readable form, it is usually just as easy to analyze every item. This is because most of the work is in setting up the programs, not in processing the items, and because the computer must read every record for the decision of whether to include or exclude the record from the sample. Aside from special cases, the main disadvantage of a census is usually the high cost relative to the other options.

Judgment Sample

Judgment sampling is not statistical or scientific sampling: it is discretionary. In this type of sampling, the evaluator bases the selection of a sample on knowledge or judgment about the characteristics of the population. Haphazard or "catch as catch can" samples—for example, grabbing a few items "at random"—are usually included in the category of judgment sampling.

Judgment samples have valid uses. When one need not generalize to a population, a census or a statistical sample is not necessary, and a small judgment sample might be cost effective. For example, if the objective of an evaluation is to show vulnerability to fraud (without regard for the probable incidence of fraud), a judgment sample may be satisfactory.

The case study approach uses judgment sampling. By definition, one of the features of the case study strategy is that it is not a census and does not involve statistical sampling of the primary sampling units and therefore one cannot generalize to the population. There are a variety of situations in which case studies, and thus judgment sampling, would be appropriate. (See Case Study Evaluations, transfer paper 10.1.9, issued in November 1990.)

Sometimes the job objective is to generalize, but it is not possible to obtain a sampling frame (either a listing of the population or a rule for determining the population). Statistical sampling is then not possible, and we may be forced to use a judgment sample. The key problem with using a judgment sample when we want to generalize to the population is that we have no way of knowing how near the sample results are to the population characteristics we are attempting to estimate. Although not necessarily less accurate than probability samples in describing a population, judgment samples lack three characteristics of statistical samples: (1) random selection of the units to be examined, (2) mathematical determination of the sample size, and (3) mathematical measurement of the risk of being wrong because a sample was used. (The precision or sampling error can be calculated objectively for any level of confidence or assurance desired, and it can be stated within the selected confidence or probability that the sample result will not vary from the true but unknown population parameter by more than the calculated precision or sampling error.)

Statistical Sample

When the objective of an evaluation is to draw conclusions about a population of people or things and when we can develop a sampling frame, statistical sampling is the method of choice. Because no individual's judgment is infallible and because the ability to make effective judgments varies widely from individual to individual and even in the same individual from time to time, the evaluator's judgment and objectivity can always be questioned when using judgment sampling. This is not so in statistical sampling, which is based on the widely accepted theory of probability, because the sample is scientifically selected. Certainly, the complaint that the evaluators looked at only the worst cases would have no merit.

Using statistical sampling, another party can repeat a study and expect to reach the same numerical conclusions about the characteristics of the population being measured. Although the study results may be interpreted differently, there can be no question about the numerical calculations. Likewise, statistical samples can be combined and evaluated even if they were taken by different persons. Evaluators working at different locations can participate independently in the same job, and the results from several locations can be combined to develop one estimate of the population parameter. Also, a study started by one evaluator can be continued by another without difficulty. Further, if evaluators decide to extend the sampling, they can do so easily and combine the results.

Statistical sampling provides a means of objectively determining the sample size necessary to provide sample results having a certain measure of the risk of being wrong required for the population being examined and the evaluation question being answered. This approach usually results in a smaller sample, with resultant savings in time and money, than that found in using judgment sampling. Because of the intuitive but incorrect belief that an adequate sample must always be a fixed percentage, say 5 or 10 percent, of the population, oversampling occurs frequently. For example, if the population were 131,000 and the percentage chosen were 5 percent, then the sample size would be 6,500, which is larger than necessary for probability sampling. However, if the population were small, using the intuitive approach of selecting a sample size that is equal to a fixed percentage of the population could yield a sample too small to produce sample results that have measures of the risk of being wrong that are acceptable for the particular job. For example, if the population consisted of 200 items and a 10-percent sample were drawn, the sample size would be only 20 items.

Statistical sampling may sometimes be a powerful method of discovering fraud or misuse of resources. After several reviews, an agency employee might be able to figure out the evaluators' selection pattern if they used judgment sampling. The employee could then arrange the files so that the evaluators could not select documents containing evidence of fraud. In probability sampling, all the documents in the population have a certain nonzero probability of selection, and manipulating their location will not affect this probability. Also, because statistical sampling results in the evaluator's looking at items that can occur anywhere in the population, agency employees may believe that the evaluators are making a more thorough examination and therefore may be less likely to continue the fraud or other abuse.

A particular project may use a combination of sample approaches. For example, runaway-youth centers might be chosen judgmentally, but within each center information could be sought from a random sample of youths who use a center's services. The most appropriate combination depends upon the job's objectives and constraints.

Determining the Type of Statistical Sampling

If a statistical sample is the choice, a further decision must be made among the possible types of statistical sampling methods. Among the types that might be used, three common ones—simple random sampling, stratified sampling, and cluster sampling—are discussed later in this chapter. Two additional sampling types, discovery sampling and acceptance sampling, are relevant to some evaluations as described in chapter 6.

Determining the Sample Size

The determination of an appropriate sample size is part of sample design. However, we do not treat sample size in this discussion of sample design for two reasons: (1) several of the factors that must be considered in the calculation of sample size, confidence level and precision, are not introduced until the next chapter and (2) sample size depends also on the estimation procedures discussed in chapters 4 and 5.

To use this paper for guidance in determining sample size, evaluators should decide on the sampling method to be used, the estimation procedure, the confidence level desired, and the precision that is required to meet the objectives of the job. Reference to the appropriate sections on calculating sample size will then provide the necessary guidance.

Selection Procedures

Selection procedures involve the method of actually picking the sampling units (sometimes called drawing the sample). All types of statistical samples use random selection procedures. The selection procedure may be dictated by the population's arrangement, the evaluator's knowledge or "guesstimate" about how the sampling units are arranged within the population, the proportion of the population that will be drawn into the sample, or the method used to identify the sampling unit. For example, if the sampling units were to be tax returns that the evaluator knew or guessed were stored in bundles of 100 but that were not ordered in any fashion, then the sampling selection method might be a two-part random selection procedure. That is, the first random selection picks the bundle and the second selects the return within the bundle.

Practical selection procedures are discussed in detail in chapter 7, but a short example will illustrate one procedure. Consider the runaway-youth program example again. Suppose we wish to use the simple random sampling design for selecting 50 center directors from a population of 200. One procedure would be to write the name of one director on one of 200 Ping-Pong balls, one for each center in the population, and to put the 200 balls into a jar. The jar would be thoroughly shaken, and a person would draw 50 balls from it to form the sample. In this procedure, each ball, and therefore each director, would have an equally probable chance of being included in the sample. The procedure would be random, and the sample would conform to the requirements of the simple random sampling design.

Estimation Procedures

Estimation procedures, discussed in chapters 4 and 5, refer to the mathematical formulas used to calculate both the estimates of the population characteristics of interest obtained from sampling and the precision of these estimates. For example, we might take a sample of Farmers Home Loan Administration loans and categorize them according to their loan status. From the results of our sample, we might say that 33 percent of the loans issued by this agency were in default and we are 95-percent confident that the true but unknown percentage of defaulted loans is somewhere in the range between 28 percent and 38 percent. Another way of stating this estimate would be that we are 95-percent confident that the true but unknown percentage is 33 percent plus or minus 5 percent. The level of confidence tells us how much confidence we have in our estimate, and by subtracting this level from 1, we get the risk of being wrong. The precision (the plus or minus 5 percent in the example above) provides the range in which we feel confident that the true population characteristic actually lies. The various types of computation methods, such as manual calculations, with or without a calculator, or computer calculations, may be considered part of the estimation procedure.

To briefly illustrate an estimation procedure, we can consider the runaway-youth example again. Suppose we want to use the information acquired from our sample of directors to estimate the total number of staff members employed by all the centers. If simple random sampling was used, the estimation procedures are easy.

Fifty of the center directors, or one fourth of the population of center directors, were in our sample. If the 50 directors reported, collectively, that 287 staff members worked in their centers, than our best estimate of the total staff members for all centers would be 4 times 287, or 1,148. In this simple example, the population estimate is just inversely proportional to the sampling fraction of one fourth.

The population estimate above will almost certainly be incorrect by some amount because we sampled and did not do a census. However, by using the concepts of confidence level and precision, we can also estimate the amount of uncertainty in our estimate. Procedures for doing so are discussed in the next two chapters.

Data collection from a sample seldom proceeds exactly as planned. When we get nonresponses to questionnaires or when the respondent does not answer all the questions or the like, special estimation techniques are required. Some of the interplay between sampling and data collection problems is discussed in chapter 8. In general, evaluators should consult with a specialist for advice on how to cope with data problems when making estimates.

Three Statistical Sampling Methods

Simple Random Sampling

Simple random sampling is the simplest method of drawing a statistical sample, and this design is the basis of all the other sampling designs. The assumptions underlying the use of simple random sampling are that the population is similar—that is, there is only moderate variation among the values of the items in the population—and is in one location, or it can be sampled from a single list of sampling units if it is in several locations. Once the population list has been developed, the sample can be drawn by using one of the selection procedures described in chapter 7 or appendix II. No attempt is made to segregate or separate any portion of the population into separate groups before the sample is selected. Thus, all individual items in the population have an equal probability of being included in the sample. This is the most commonly described method of sampling but is sometimes less efficient than other methods.

An example of simple random sampling in GAO work is selecting a random sample of children participating in one school district's lunch program, in order to determine the proportion of children whose family income meets the program's eligibility criteria. We would get a list of all students in the school district's lunch program and select a random sample of x students and determine in some fashion whether or not the family income met the eligibility criteria.

Stratified Sampling

Stratified sampling refers to the situation in which the population is divided into two or more parts (strata) and a simple random sample is selected for each part (stratum). An estimate is determined separately for each stratum, and these are combined to form an estimate for the entire population. A stratum is a subpopulation from the total population. The terms "high income," "middle income," and "low income" indicate three strata of a population of people classified by the income they received. Tax returns might be divided into strata based on the asset size of the company submitting the return—for example, under $250,000, $250,001 to $500,000, $500,001 to $1,000,000, $1,000,001 to $5,000,000, and so on. A stratified sample can be used to

- obtain equal precision with a smaller sample or tighter precision with the same size sample. Stratification generally reduces the cost of a sample for a given precision;
- obtain separate estimates for the groups in the individual strata, if such estimates would be useful for comparison purposes; and

- give special emphasis to certain groups within the population, such as invoices of high dollar values or those with a great error potential.

Sometimes stratification is necessary because the population is divided up among several locations and it is not possible to develop a single sampling frame. For example, the objective of the job may be to estimate the dollar loss to the government because of the errors in tax returns filed by companies. If it were not possible to develop a single list of companies across the offices where the tax returns were filed and stored, a separate sample would have to be drawn at each office, and estimates for each office would have to be combined in order to obtain one overall estimate for the entire population.

Stratification may be desirable if the costs of data collection differ from stratum to stratum. For example, in one stratum we may have to collect the data by personal interview but in another stratum we may be able to use mailed questionnaires.

When defining strata and setting their boundaries, evaluators should keep certain rules in mind. (1) Each sampling unit can be included in one, and only one, stratum. (2) The strata must not overlap. (3) The sampling units in each stratum should be as much alike as possible in relation to the characteristic being measured.

Each stratum is treated as if it were a separate population from which items are selected independently; that is, the sample selected in one stratum must not depend on, or be related to, the sample selected in another stratum. One of the acceptable random selection procedures is used to draw the sample in each stratum.

The total sample may be allocated to each stratum in proportion or in disproportion to the number of sampling units in that stratum. With proportional allocation, the sampling fraction (sample size divided by the population size) is the same in each stratum. With disproportional allocation, sampling fractions differ in two or more strata. Disproportional allocation may be based on professional judgment or on mathematical formulas, in order to minimize the overall precision or the overall cost of data collection. Appendix III discusses the allocation of sample size to strata as well as the calculation of estimates, precision, and sample sizes with a stratified sample design.

Cluster Sampling

Another type of sampling is cluster sampling, which is the selection of groups of sampling units (or clusters) rather than the selection of individual sampling units directly. Examples of clusters are folders in filing cabinet drawers, baskets of produce, counties in a state, and the persons in a household. We are sometimes able to examine all the sampling units within the sampled cluster. However, if the clusters are large, it is often preferable to select a random sample of units within the selected cluster. This is referred to as two-stage cluster sampling.

Because of the size and complexity of some populations, cluster sampling must on occasion be done in more than two stages as described above. The technique is very similar to the two-stage but is extended. For example, if three stages are needed, we would first take a sample of clusters (called primary sampling units). Then we would take a sample of units from within the cluster (called secondary sampling units). Finally, we would take a sample of the elements or the unit or object or thing or person on which we take our measurement from each of the selected secondary sampling units. An example of cluster sampling is given in appendix III.

Basic Estimation Procedures and Further Sample Design Considerations

Almost invariably, one of the first questions someone interested in sampling asks is, "How large a sample do I need to take?" Procedures for calculating sample size are presented in this chapter. It also introduces the concepts of precision and confidence level. In the context of sample design, these concepts lead to a determination of the sample size appropriate for a particular evaluation.

Basic estimation procedures are presented for two situations: variable sampling (means and totals) and attribute sampling (proportions and number of occurrences).[1] Variable sampling is used when we are estimating something that can be quantified (that is, the measurement is on either the interval or ratio scale). This measurement is known as a variable. Some examples of these continuous variables that we have used in GAO are (1) the dollars of interest that the Internal Revenue Service owed a taxpayer, (2) the dollars it would cost to bring a railroad crossing up to current safety standards, and (3) the number of employees for a company.

When sampling for attributes, we want to determine what percentage or proportion of the population has the characteristic we are interested in (that is, we are using the nominal or ordinal scale of measurement). Either the sampling unit has the characteristic or it does not, although a third unknown value can be ascribed. Sometimes the characteristic is only one of several choices, as when people are classified by gender, race, educational level, or employment status. Examples of the characteristics for which GAO might sample are unapproved travel orders, health

[1]These terms are defined in the glossary and have different meanings to the different disciplines. "Variable sampling" is sometimes called sampling for variables; similarly, "attribute sampling" is sometimes called sampling for attributes.

insurance claims that were paid without supporting documentation, and farm loans that are unpaid.

We can use a single sample to develop estimates for proportions and for means and totals. For example, in examining health insurance claims, we can use one sample to estimate both the proportion of occurrences of undocumented payments and the dollar amount of those payments. However, in general, since the variation in variables is larger than in proportions, the sample size required for estimating means and totals is larger than that for estimating proportions. Therefore, when we calculate the sample size, we should use a calculation based on the most important estimate for the objective of the job.

The Concepts of Precision and Confidence

Specifying the precision needed for sample estimates is an important part of sample design. The desired precision is the amount of sampling error that can be tolerated but that will still permit the results to be useful. This is sometimes called tolerable error or the bound on error.

Because precision is a way of expressing the amount of error that can be tolerated, it is related to the accounting concept of materiality or the evaluative concept of importance. The notion of materiality, according to a 1957 statement of the American Accounting Association, says that an item should be regarded as material if there is a reason to believe that knowledge of it would influence an informed investor's decision. In policy or evaluation research, a result is considered important if there is a reason to believe that knowledge of it would influence a decisionmaker's behavior or be important in public debate.

Materiality, or importance, is a relative concept rather than absolute. For example, a $100,000 overstatement of the assets of a company whose total assets are only $200,000 may be material. A $100,000 overstatement of the total assets of a multibillion dollar corporation would probably be immaterial. A 10-percent misstatement about the notes receivable account of a small loan company would probably be material, but a 10-percent misstatement in the office supplies account balance would probably be immaterial. Since importance and materiality are relative, we need a basis for establishing whether a finding is important or material.

Materiality, or importance, is linked to precision in the following way. To develop a reasonable specification of precision, evaluators must gauge the materiality or importance of the estimates to be made and use this information to decide how much the statistical estimates can vary from the true but unknown population value and yet provide useful information. Going back to the example above, if we were attempting to verify the notes-receivable balance, we would probably be very unwilling to allow the estimate to vary from the actual amount by as much as 10 percent. Thus, we would probably want to take a large enough sample of individual loans and confirm the balances to keep the estimate well within 10 percent of the true but unknown population value. However, if we were evaluating an account that was a small amount of the total that we were evaluating, we could probably live with a misstatement of 50 to 60 percent.

In addition to specifying the precision of the estimate, evaluators must specify the degree of confidence that they want placed in the estimate. Referred to as confidence level, this is expressed as a percentage. It is the complement of the chance that our estimate and its precision will not contain the true but unknown population value. (The concept of confidence is developed more fully in appendix I.) The confidence level should be determined by the importance of the sample results to the overall objectives of the job.

Evaluators should decide on the confidence level during the design phase shortly after defining the objective. The decision should not be postponed until after a sample has been taken and evaluated, in order to get a confidence level that makes the precision or sampling error look smaller. Although the point estimate is our "best guess" of the true but unknown population value, the chance of its being 100-percent correct is infinitesimally small. Half the time, the "true" value is larger than our point estimate; half the time, smaller. The point estimate is almost always considerably different from the "true" value. The essence of a statistical estimate is, therefore, a statement of probability that the "true" population value (that is, what a 100-percent examination would disclose) is between two stated values. For example,

- We are 95-percent confident that the error in the population is between 3.1 percent and 4.3 percent.
- We are 95-percent confident that the error in the population is 3.7 percent plus or minus 0.5 percent.
- We estimate that the population total is $21,400, and we are 95-percent confident that the true but unknown total lies between $20,100 and $22,700.
- We estimate that the population total is $21,400, and we are 95-percent confident that the true but unknown total is $21,400 plus or minus $1,300.

Other, related considerations are the costs and time required to obtain the sample data. If the precision is specified "too tight," without considering cost and time, the sample size may be larger than practical. Usually, only limited resources (of money, staff, and so on) and time are available for data collection. This must always be remembered when the estimate's precision is being specified.

Various mathematical formulas can take data collection costs into account when we compute sample sizes and specify precision. However, such formulas are beyond the scope of this paper. Perhaps the practical guidance that can be given here is that the evaluator should first specify the desired precision to meet the objective and then, with the assistance of the specialist, calculate the sample size needed to achieve this precision and finally estimate the cost or time required to collect the data for the computed sample size. If the cost is more than can be afforded, the precision should be relaxed (allowed to be larger) until an affordable sample size is found, or the evaluator should decide that the job cannot be done since the resources do not allow a sample size to provide a precision that is acceptable. An adjustment like this should be made by relaxing the specified precision, not by manipulating the confidence level.

Variable Sampling

As we discussed above, a variable is something that we measure on the interval or ratio scale—dollars, the count of items in an inventory, and the like. When we are sampling for variables, we usually want to estimate the total value for the population of interest—for example, the total amount of assessed taxes that were not collected. For some jobs, however, the mean may be more important to our evaluation objective.

The first step in estimating the population total is to compute the mean of the sample values using the formula in chapter 2. The mean is simply the sum of the sample values divided by the sample size. The mean, a very important measure of central tendency, can be manipulated mathematically, which is not true of the other measures of central tendency that we mentioned (median and mode).

For certain types of data, the median is a better measure of central tendency than the mean. The median, the middle value in a set of values, is selected in a way such that half the values are below it and half are above. Thus, the median is a locational measure of central tendency. An example of using a median in GAO work appeared in a report on the length of sentences that people who used handguns in committing their crimes received. Since we could not place a value in terms of years on a life sentence, the report described the length of confinement in terms of the median sentence.

An extremely simplified instance of variable sampling might occur if our objective were to estimate the dollar amount of small purchases made by an agency during a specific fiscal year. The population would then be defined as all small purchases during the fiscal year. During the year, there were 100 such purchases. (It is somewhat unusual to sample from such a small population, although doing so may be necessary on occasion; we use the small population here for its convenience as an example.) Now examine table 4.1.

Table 4.1: Example of Sample Values

Item	Amount
1	$147
2	259
3	185
4	164
5	150
6	187
7	137
8	159
9	125
10	172
11	277
12	142
13	231
14	125
15	172
16	241
17	232
18	233
19	205
20	226
21	236
22	202
23	89
24	248
25	160
26	194
27	177
28	135
29	96
30	163
Total	**$5,469**

Using the formula in chapter 2 for calculating the mean, we calculate from the data in table 4.1 that the mean is 5,469—that is, the total value for the 30 items in our sample—divided by 30, or 182.30. Thus, we could say that the average purchase for this agency during the fiscal year was $182.30.

In most GAO evaluations, as noted above, we are more interested in the population total than the mean. To estimate the population total, we simply assume that the sample mean is an estimate of the population mean and multiply the sample mean by the number of elements in the population. This is called expansion or extension estimation. For the example above, we would then estimate the total by multiplying the mean of $182.30 by the 100 elements in the population (100 x 182.30 = 18,230). Thus, the estimated total amount of small purchases made by the agency during the fiscal year in question is $18,230.

Does this adequately estimate the true but unknown total for the population? How can we be sure? This depends on how good our assumption was that the sample mean is an estimate of the population mean. To measure how good our assumption is and therefore how good is our estimate of the total, we have to determine the precision of the estimate and the confidence level at which the precision is stated.

To compute the precision of the estimate (at this point, precision is not the desired precision but the actual precision from the results of the sample), or sampling error, of the estimated total, we first compute the standard deviation of the purchase amounts. The standard deviation is a numerical measure of the spread of a group of values about their mean. Understanding this statistic is a key to understanding much of sampling. As we saw in chapter 2, the standard deviation is a measure of the average squared deviation from the mean. The first step is to get the deviation (or difference) of each item from the mean. These items are first squared and then summed. This result is then divided by the sample size minus 1.[2] Finally, the square root is taken. Engineers call this statistic the root mean square, because it is the square root of a form of the average of the squared deviations. This statistic is always in the same unit of measurement as the element itself. For example, the standard deviation of the example above will be in dollars since the sampling units were measured in dollars.

The standard deviation (S) can be defined by the formula

$$S = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$

By doing the calculations either by calculator or by computer for our example, we would find that the sample standard deviation is $48.71.

The next step is to calculate the precision or sampling error of the mean (SE) at a specified level of confidence. This is done by multiplying the standard deviation by a "t" value corresponding to the stipulated level of confidence and dividing by the square root of the sample size. (Appendix I contains a table of "t" factors for commonly used confidence levels.) The formula is

$$SE = \frac{ts}{\sqrt{n}}$$

[2]Note that we divide by n – 1 rather than the full sample size n. Stated simply, the reason for doing this is that we have used the sample mean to estimate the population mean, which we do not know. The effect of this is to "use up" one of the sample values, leaving only n – 1 values as a basis for estimating the standard deviation. We lose one value (technically, one degree of freedom) for every population value, such as the mean, that we estimate from the sample.

Suppose that we have previously decided that the confidence level for the precision of our estimate should be 95 percent. The "t" factor for 95 percent is 1.96 (or 2 for all practical purposes). Using this value as well as the others, we obtain the precision or sampling error at 95-percent confidence by

$$(1.96)(48.71)/\sqrt{30} = 17.43$$

Thus, the sampling error of the mean is $17.43 at the 95-percent level of confidence.

We mentioned earlier that in most GAO sampling applications, we are interested in the estimated total. How do we compute the sampling error of the total? We computed the estimated total by multiplying the sample mean by the number of items in the population. The computation of the sampling error of the total is a parallel procedure. We simply multiply the sampling error of the mean by the number of items in the population. For our example, we obtain 100 x 17.43 = 1,743. Thus, the sampling error of the total at the 95-percent level of confidence is $1,743.

The interpretation of this value parallels the interpretation of the sampling error of the mean. For practical purposes, using the 95-percent confidence level, we state that if all small purchase orders were reviewed in the same fashion as the sampled items, the chances are 19 in 20 that the estimate obtained from the sample would differ from the true but unknown population total by less than the sampling error. Note that sampling errors are always stated with an associated confidence level. The point estimate, the point that is likely to be closest to the true population total, is $18,230.

When we sample in GAO, a single sample may be used to develop estimates for many different population values. In principle, if different values are to be estimated from the same sample, the sampling errors should be made larger to account for the fact that each estimate has a 5-percent risk of being wrong, and so we have increased our exposure to making a bad estimate. (See Dixon and Massey, 1983, and Snedecor and Cochran, 1989.)

Calculating Sample Size

Whether we are sampling for variables or for attributes, one advantage of statistical sampling is that it permits us to determine objectively the sample size required to achieve a given degree of precision at a specified confidence level. To make this computation, we need to estimate the standard deviation of the population. It may seem paradoxical that we must obtain information about the population when we are sampling in order to estimate its characteristics. However, when we look at other types of information-gathering, this is really not so strange. To look up a word's correct spelling in a dictionary, we must have some idea of how the word is spelled. To locate our position on a map, we must know "about where we are." And to compute a vessel's exact position by celestial navigation, we use an assumed position that has to be a fairly accurate estimate.

In computing the sample size, we must consider three factors. The evaluators specify two factors—the confidence level and the desired precision of the estimate. The third factor, the standard deviation, is based on the characteristics of the population. A formula can bring these three factors together for computing the sample size. (The population size is not taken into account, as we will explain below.)

To illustrate the sample size formula, let us assume that the sample of 30 purchases given above was the result of a sample taken during the scoping phase of the job. Also let us suppose that the evaluator wanted to reduce the precision to $1,400 from the $1,743 at the 95-percent confidence level. The first step is to convert the desired precision (or tolerable error) of the total to the tolerable error of the mean. We do that by dividing the desired precision of the total by the population size, which in this example is $1,400 divided by 100, or $14.00.

Once we have the tolerable error of the mean, we compute the required sample by using the formula

$$n = \frac{t^2 s^2}{E^2}$$

where E is tolerable error. In our example, since we want a 95-percent confidence level, we use a "t" value of 2, the standard deviation that we calculated from our initial sample, and the tolerable error of the mean of $14.00 in the formula. Thus we have 2 times 48.71 divided by 14, which gives a sample size of 49. This means that an additional 19 purchase orders, in addition to the first sample of 30, would have to be sampled in order to achieve the required precision.

Notice that we have used the standard deviation obtained from our first sample as an estimate of the true population standard deviation. If the true standard deviation were known, we would not have to sample, since we would know the mean and standard deviation.

The best method of estimating the standard deviation is to take a small, random, preliminary sample and calculate the standard deviation from it. The sample should be random so that, if it must be increased to obtain the desired precision, which usually happens, the preliminary sample can be included in the final sample and no work will have been wasted. The preliminary sample should consist of at least 30 cases; otherwise, the statistical laws discussed in appendix I will not apply or will not work as well.

Sometimes it is possible to use the results of samples taken at other times as an estimate of the standard deviation. This is usually satisfactory if a similar review has been made and no major change in the distribution of the population values is suspected. Another possibility is that subject matter experts may be able to guess the size of the standard deviation from their knowledge of the field or previous work.

Occasionally, it is stated that a larger population requires a larger sample or that the sample size must always be a certain percentage of the population. This is incorrect. As we noted in the formula for calculating sample size, the size of the population does not enter into the calculations. The population size and the sample size are slightly related, but before explaining this further, we need to discuss the concepts of sampling with replacement and sampling without replacement.

When we sample with replacement, an item selected for the sample is returned to the population and can be selected again. Since the sample item is replaced, the population from which the sample is drawn can be regarded as infinite. (In theory, when we sample with replacement, the entire sample could consist of the same item.) When we sample without replacement, an item selected for the sample is "used up" and cannot be selected again. Thus, each item can appear in the sample only once.

Sampling without replacement is used in GAO, except in special circumstances. Because we are gradually using up the population, we would expect that our estimates would be better as we go along. In fact, if our sample were 100 percent of the population, we would have an exact estimate of the population mean. If the sample size is large in relation to the population size, we can use this efficiency to reduce both the sample and the sampling error. We do this through the finite population correction (FPC) factor.

Considering a practical matter, we need use the FPC only when the sample size is greater than 5 percent of the population.

To use the FPC to reduce the sampling error of the mean (and, by extension, the sampling error of the total), we multiply the sampling error by the factor

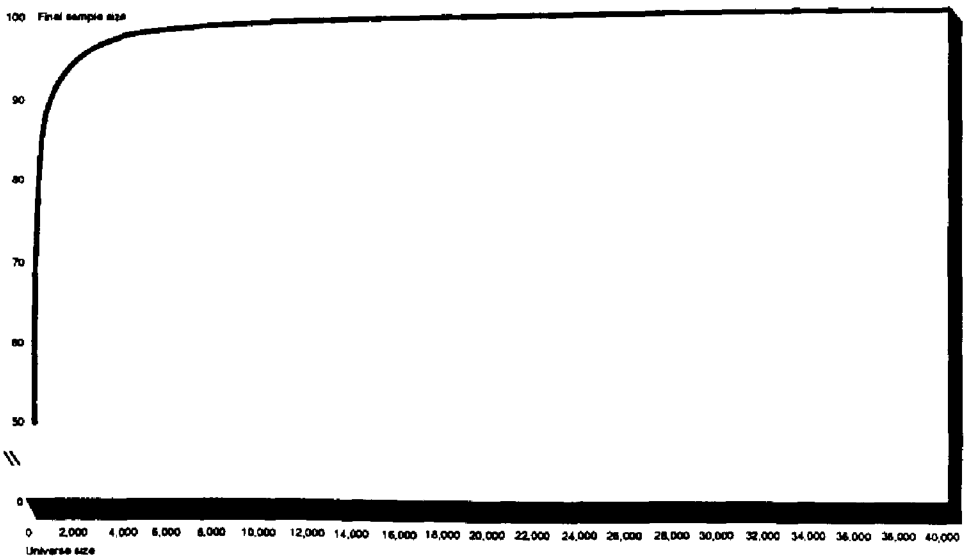
$$\sqrt{\frac{N-n}{N}}$$

Using the data from our random sample of 30 purchase orders drawn from the population of 100 purchase orders, we multiply our sampling error of the mean $17.43 by the square root of 0.7, which is the population size minus the sample size all divided by the population size. Thus, the adjusted-for-population-size sampling error of the mean is $14.58 and, thus, the adjusted-for-population-size sampling error of the total is $1,458. By using the FPC, we have in this case reduced our sampling by about 16 percent ($1,458 versus $1,743 without the FPC).

To use the FPC to reduce the sample size, we first calculate the sample size. If it is greater than 5 percent of the population, we enter this first estimate of the sample into the following calculation to determine the adjusted sample size. In our example above, we calculated that a sample size of 49 purchase orders would be required in order to reduce the sampling error of the total to $1,400. Since 49 purchase orders obviously make up more than 5 percent of the population, we calculate the adjusted sample size by dividing 49 by the sum of 1 plus 49 divided by 100, or we have 49/1.49, which equals 33. Thus, the adjusted sample size is 33 purchase orders. By using the FPC, we have reduced the required sample size from 49 to 33. If the FPC is used to compute the required sample size, it should be used to compute the sampling error. Otherwise, the sampling error will be greater than the desired precision specified. If the sample size is 5 percent of the population or less, the FPC may be ignored.

Figure 4.1 highlights the weak relationship between population size and sample size, especially when the population is large compared to the sample size. The figure shows the adjusted sample size for various population sizes, assuming that the initial estimate of each sample size was 100. Table 4.2 provides the data for the figure.

Figure 4.1: Final Sample Size As Population Size Increases[a]



[a]First estimate of sample size is 100.

Table 4.2: Data for Figure 4.1

Population size	Adjusted sample size
100	50
200	67
400	80
600	86
800	89
1,000	91
2,000	95
4,000	98
6,000	98
8,000	99
10,000	99
20,000	100
40,000	100

Note that if the first estimate of the sample is less than 1 percent of the population size, the adjusted sample size equals the first estimate and remains constant.

Attribute Sampling

Sometimes we want to estimate the proportion, percentage, or total number of items in the population that possess some characteristic or attribute or that fall into some defined classification. Examples are the percentage of the labor force that is unemployed, the percentage of people older than 65, and the number of low-income households in a county.

Assume that the evaluators are reviewing a supply depot's efficiency of operations. For this evaluation, they want to estimate the number of requisitions that the depot was unable to fill during the past fiscal year because requisitioned items were out of stock and the rate at which the depot was unable to fill them. The population consists of all 12,000 requisitions received by the depot during the fiscal year.

The characteristic of interest is, of course, a requisition that was not filled because the item was out of stock. For this example, assume that the sample size is 100 and that the evaluators found in the 100 selected requisitions 36 that were unfilled because the item was out of stock.

One of GAO's tools is a statistical package called SRO-STATS, used in the Training Institute's course on statistics for evaluators. The input of these sample data into the program for attribute sampling produces the output shown in figure 4.2.

Figure 4.2: Sample Data for Supply Depot Example

```
INPUT DATA:

     POPULATION SIZE = 12000

     SAMPLE SIZE =          100

     NUMBER OF ATTRIBUTE OCCURRENCES =        36


SAMPLE STATISTICS:

     PROPORTION OF OCCURRENCES =0.3600

     EST. STANDARD ERROR OF THE PROPORTION =0.0480

     EST. RELATIVE ERROR OF THE PROPORTION =0.1335

     EST. 95% CONFIDENCE INTERVAL =   0.3600 +&-   0.1003

POPULATION ESTIMATES

     EST. TOTAL OCCURRENCES IN POPULATION =4320

     EST. STANDARD ERROR OF TOTAL OCCURRENCES 577

     EST. 95% CONFIDENCE INTERVAL =              4320 +&-1204
```

To get an estimate of the total number of occurrences in the population, we compute a rate of occurrence (analogous to the mean) from the sample data, assume that it is an estimate of the rate of occurrence in the population, and then multiply it by the population size.

Given an estimated rate of occurrence of 36 percent for unfilled requisitions in our population of 12,000, after multiplying the population size times the rate of occurrence, we estimate that the number of unfilled requisitions is 4,320. Do these estimates of the rate of occurrence and the number of unfilled requisitions actually represent the true values in the population? Can we calculate their precision? Yes. The laws governing large samples discussed in appendix I for variables apply to attributes estimation.

From the results obtained above, we can say that the number of unfilled requisitions at the 95-percent confidence level is within a range of 1,204 on either side of 4,320. That is, the number of unfilled requisitions falls between 3,116 and 5,524 at the 95-percent confidence level, or the best estimate of the number of unfilled requisitions is 4,320 with a sampling error of 1,204 as stated at the 95-percent confidence level.

It is worth repeating that although the point estimate is our "best guess" of what the "true but unknown" population characteristic is, the chance of its being 100-percent correct is infinitesimally small. Half the time, our point estimate is larger than the "true" value; half the time, smaller. The point estimate is almost always considerably different from the "true" value. The essence of a statistical estimate is, therefore, a statement of the probability ("confidence" or "certainty") that our calculated confidence interval contains the true but unknown population value.

In these computations, we have used the rate of occurrence found in the sample to represent the unknown rate of occurrence in the population. If we increase the sample size to 400 items, the estimated rate of occurrence of unfilled requisitions should be about the same, but the sampling at the 95-percent confidence level would be reduced to 4.8 percent, or 576 requisitions.

The unknown percentage we are trying to estimate is fixed, a constant; it does not move around. Only the estimates, both point and interval, from different statistical samples vary.

If a large number of samples of the same size were taken from the same population, 68 percent of them would have their individual point estimates within 1 standard error of the "true but unknown" population percentage, about 95 percent would be within 2 standard errors, and 99 percent would be within 2.58 standard errors. At the 95-percent confidence level, we could state (rather crudely) that if all 12,000 requisitions in the population were examined in the same fashion as the sample items, the chances are 19 out of 20 that the results would differ from the estimate obtained from the sample by less than the sampling error.

The concepts and formula used in calculating sample sizes for attribute sampling are the same as those for variable sampling. The SRO-STATS program that we ran above has as one of its options the calculation of the necessary sample size for a given precision. (See figure 4.3.)

Figure 4.3: Sample Data for Supply Depot Example for a Given Precision

```
INPUT DATA:

     POPULATION SIZE = 12000

     SAMPLE SIZE =          529

     NUMBER OF ATTRIBUTE OCCURRENCES =        190


SAMPLE STATISTICS:

     PROPORTION OF OCCURRENCES =0.3593

     EST. STANDARD ERROR OF THE PROPORTION =0.0204

     EST. RELATIVE ERROR OF THE PROPORTION =0.0568

     EST. 95% CONFIDENCE INTERVAL =    0.3593 +&-   0.0414

POPULATION ESTIMATES

     EST. TOTAL OCCURRENCES IN POPULATION =4311

     EST. STANDARD ERROR OF TOTAL OCCURRENCES 245

     EST. 95% CONFIDENCE INTERVAL =                4311 +&- 496
```

Thus, we would have to sample 429 requisitions in addition to the 100 already sampled.

To use the sample size formula, we need some "estimate" of the expected rate of occurrence of the characteristic of interest. In this example, the expected rate of occurrence came from our initial sample of 100. The expected rate of occurrence may be obtained from a preliminary random sample of at least 30 items, from prior experience in a similar review, from experts in the field, or from information supplied by the agency being evaluated. If the evaluators suspected that the agency's estimate was too low, they could increase it by 10 to 20 percentage points. This would give a larger sample size. The largest sample sizes are needed when the rate of occurrence is around 50 percent. Smaller sample sizes can be used to obtain the same precision when the estimated value moves away from 50 percent in either direction. When no other estimate is available, the sample size can be estimated with a 50-percent rate.

A Distinction Between Precision and Accuracy

We use the word "precision" rather than "accuracy" throughout this paper. Precision refers to the maximum amount, stated at a certain confidence level, that we can expect the estimate from a single sample to deviate from the results obtained by applying the same measuring procedures to all the items in the population. Accuracy refers to the difference between the value of the population from which the sample is selected and the true characteristic that we intend to measure.

We can use a simple example to illustrate the distinction. Suppose we want to estimate the mean weight of all men employed at GAO today. We could select a sample of the men, weigh them, and compute their mean weight. If the sample were large enough, we could estimate the mean weight with a very small precision. But if the list of male employees from which we drew the sample were a year old, we would not be estimating the mean weight of the male GAO employees of today; we would be estimating the mean weight of the male GAO employees of 1 year ago. The problem here is that the population from which the sample was selected is different from the population we defined as our population of interest. Thus, the estimate, regardless of how precise it might be, would be inaccurate.

Then suppose we are able to select a sample of 100 male employees from an up-to-date list of all male employees and weigh them on a single scale. If we find that the mean weight is 170 pounds with a standard deviation of 34 pounds, this estimate would have a precision of about 6.7 pounds at the 95-percent confidence level. If we wanted the estimate to be more precise, all we would do is increase the sample size. A sample of 400 would give a precision of about 3.3 pounds at the 95-percent confidence level. Or, if we weighed all the men, the result would be perfectly precise; there would be no sampling error at all, because the sample size is equal to the population size. However, if the scale were 1 pound off, and we did not know this, the results, regardless of the sample size and the degree of precision, would not be accurate. The mean weight we computed would be 1 pound less (or more) than the true but unknown mean weight. While we are concerned about bias and imprecision, statistical analysis does not provide much help in the estimation of the effects of bias on the results but it does provide a good measure of the imprecision from sampling. This is why we talk of the precision of the estimate and not the accuracy of the estimate.

Advanced Estimation Procedures

Chapter 4 discussed expansion or extension estimation, in which we calculate a sample mean or proportion and multiply it by the population size to obtain the estimated total or the estimated number of occurrences in the population. In this chapter, we describe ratio, regression, and difference estimation, which take into account other information that we can obtain from the sample and that we may obtain about the population. These three procedures permit evaluators to develop more information from the sample data and frequently yield more efficient (or precise) estimates. However, the possibility of using these procedures must be considered before the data are collected. Otherwise, the benefits may be lost or it may be necessary to return to the location where the data were collected.

Ratio Estimation

Sometimes, in sampling applications, the summary statistic we want to estimate is a ratio between two values, both of which can vary from sampling unit to sampling unit. For example, we may want to estimate the ratio of costs or replacement parts sold under foreign military sales agreements to the amounts received or the ratio of Medicare reimbursements for prescription drugs to total reimbursements.

In other applications, we may want to estimate the total value of an unknown variable that is related to another variable for which we already know the population total value. For example, we may want to estimate the total subsistence cost claimed on an agency's travel vouchers for a year, when we already know the total amount of travel reimbursement (the population total) for the year and the number of vouchers paid (the population size).

In this case, we select a random sample of vouchers and record, for each selected voucher, the amount paid, which corresponds to the population total we already know and is referred to as the auxiliary variable, and the amount of the claimed subsistence cost, which corresponds to the total we want to estimate and is referred to as the primary variable. Thus, we record two values for each sample travel voucher. The summary statistic is the ratio in which the total of the primary variable (the subsistence cost) is the numerator and the total of the auxiliary variable (the amount paid) is the denominator.

We should also use ratio estimation when we suspect that there is a positive correlation between the two variables, even if we are interested not in the ratio but only in the estimated total of the primary variable. If the correlation is positive and if it is strong enough, the estimate that can be obtained will be more precise than the estimate that can be obtained with simple expansion estimation.

Ratio estimation can be used with more complex sample designs such as stratified samples and cluster samples. However, the use of ratio estimation with complex sample designs is beyond the scope of this paper.

Regression Estimation

Like ratio estimation, regression estimation is an attempt to increase precision by using additional information that we know about the population and can obtain from our sample. We obtain two measurements, one of the primary variable and the other of the auxiliary variable, on a single sampling unit. (The primary variable and auxiliary variable are defined exactly the same as they were with ratio estimation.) In this technique, we use the regression model, which is well known from data analysis, to make statistical estimates.

Like ratio estimation, regression estimation can be used with more complex sample designs.

Difference Estimation

Difference estimation is used when we want to obtain a "corrected" estimate of a previously stated "book" value. For example, suppose we wanted to estimate the correct total value of an inventory when we know the value according to the agency's records and can take a sample from the inventory items and correct the items examined in the sample, if necessary. It is also an attempt to increase precision by obtaining two measurements on a single sampling unit. However, difference estimation will increase precision only if the spread of the differences between the primary and auxiliary variables is very small.

Example of Advanced Estimation Procedures

Suppose we have a population of 10,000 small purchases from agency records for a fiscal year. We know from the records that the total amount paid for these small purchases was $5,100,000. Let us assume that we selected a sample of 50 of these small purchases and recorded the value of the small purchase as recorded on the agency records and the value that, based on GAO audit work, we say the agency should have paid. Table 5.1 shows the results of the sample of 50 items.

Using the matched data set program from the SRO-STATS package and inputting the data for this example, we would have the computer output shown in figure 5.1.

Table 5.1: Data for Advanced Estimation Procedures Example

Item	Book value	Audited value
1	$300	$267
2	900	774
3	300	255
4	200	174
5	900	810
6	700	560
7	1,000	820
8	100	80
9	900	765
10	700	630
11	700	630
12	400	332
13	300	255
14	100	84
15	200	168
16	100	88
17	600	528
18	400	340
19	900	747
20	1,000	800
21	1,000	862
22	600	504
23	800	648
24	200	176
25	200	172
26	1,000	890
27	900	792
28	600	540
29	500	525
30	200	172
31	200	178
32	500	425

(continued)

Item	Book value	Audited value
33	200	164
34	500	420
35	500	400
36	400	324
37	200	160
38	600	540
39	500	425
40	300	264
41	900	765
42	100	84
43	100	85
44	900	810
45	300	240
46	500	415
47	500	425
48	300	237
49	500	435
50	100	86

Figure 5.1: Computer Output for Advanced Estimation Procedures Example

```
MATCHED DATA SAMPLING ANALYSIS:
***  ****

YOU HAVE A POPULATION SIZE OF                              10000
FROM WHICH YOU TOOK A SAMPLE OF SIZE                          50
THE MEAN OF ALL 10000 BOOK VALUES IS                      510.00
AND THE POPULATION TOTAL BOOK VALUE IS        5100000.00

     ***** SAMPLE VALUES *******
```

ITEM NUMBER	BOOK VALUE	AUDITED VALUE	DIFFERENCE	RATIO
******	******	********	**********	******
1	300	267	-33	0.8900
2	900	774	-126	0.8600
3	300	255	-45	0.8500
4	200	174	-26	0.8700
5	900	810	-90	0.9000
6	700	560	-140	0.8000
7	1000	820	-180	0.8200
8	100	80	-20	0.8000
9	900	765	-135	0.8500
10	700	630	-70	0.9000
11	700	630	-70	0.9000
12	400	332	-68	0.8300
13	300	255	-45	0.8500

14	100	84	-16	0.8400
15	200	168	-32	0.8400
16	100	88	-12	0.8800
17	600	528	-72	0.8800
18	400	340	-60	0.8500
19	900	747	-153	0.8300
20	1000	800	-200	0.8000
21	1000	862	-138	0.8620
22	600	504	-96	0.8400
23	800	648	-152	0.8100
24	200	176	-24	0.8800
25	200	172	-28	0.8600
26	1000	890	-110	0.8900
27	900	792	-108	0.8800
28	600	540	-60	0.9000
29	500	525	+25	1.0500
30	200	172	-28	0.8600
31	200	178	-22	0.8900
32	500	425	-75	0.8500
33	200	164	-36	0.8200
34	500	420	-80	0.8400
35	500	400	-100	0.8000
36	400	324	-76	0.8100
37	200	160	-40	0.8000
38	600	540	-60	0.9000
39	500	425	-75	0.8500

40	300	264	-36	0.8800
41	900	765	-135	0.8500
42	100	84	-16	0.8400
43	100	85	-15	0.8500
44	900	810	-90	0.9000
45	300	240	-60	0.8000
46	500	415	-85	0.8300
47	500	425	-75	0.8500
48	300	237	-63	0.7900
49	500	435	-65	0.8700
50	100	86	-14	0.8600

STATISTICAL ESTIMATES

====================

METHOD	MEAN OF AUDITED	STANDARD ERROR OF THE MEAN	POPULATION TOTAL	S.E. OF THE TOTAL
EXTENSION	425.400	36.141	4254000	361412.400
DIFFERENCE	439.400	6.807	4394000	68067.650
RATIO	437.407	3.394	4374073	33937.890
REGRESSION	437.428	3.428	4374276	34284.090

	PRECISION OF THE MEAN		PRECISION OF THE TOTAL	
	95% C.L.	90% C.L.	95% C.L.	90% C.L.
EXTENSION	72.644	60.609	726438.900	606088.600

DIFFERENCE	13.682	11.415	136816.000	114149.500
RATIO	6.822	5.691	68215.160	56913.840
REGRESSION	6.891	5.749	68911.020	57494.420

The statistics in figure 5.2 are used to determine which estimation methods are appropriate for the given data. For example, extreme skewness or too few nonzero differences will prevent the use of the more precise estimation techniques. The evaluator should consult technical staff on how to interpret this information.

Figure 5.2: Statistics for Advanced Estimation Procedures Example

```
MEAN DIFFERENCE                      =         -70.6000
RATIO                                =           0.8577
CORRELATION COEFFICIENT              =           0.9956
SLOPE OF REGRESSION                  =           0.8591
Y INTERCEPT                          =          -0.7202
NUMBER OF NON-ZERO DIFFERENCES       =               50
RELATIVE ERROR OF BOOK VALUES        =           0.0844
RELATIVE ERROR OF AUDITED VALUES     =           0.0850

ESTIMATED SKEWNESS IN POPULATION OF MEAN DIFFERENCES     =-0.0897
ESTIMATED SKEWNESS IN POPULATION OF BOOK VALUE MEANS     = 0.0445
ESTIMATED SKEWNESS IN POPULATION OF AUDITED VALUE MEANS= 0.0423
ESTIMATED SKEWNESS IN POPULATION OF MEAN RATIOS          = 0.2418
ESTIMATED KURTOSIS IN POPULATION OF BOOK VALUES          =-1.2222
```

Note the change in precision for the three estimation techniques and the drastic increase over the extension method.

Ratio Estimation

Using the information given in figure 5.2, we find that the ratio of the audited value for the small purchases is 0.8577; in other words, for every $100 that the agency spent, it should have spent only $85.77. Therefore, multiplying this ratio by the book value total, we estimate that the agency should have paid only $4,374,073 for its 10,000 small purchases. As before, this is only a point estimate, and we need to place a sampling error around this estimate at the 95-percent confidence level. Using the information from the printout, we can state that we estimate that the total amount that the agency should have paid is $4,374,073, and we are 95-percent confident that the true but unknown total is $4,374,073 plus or minus $68,215.16.

Regression Estimation

Note that with regression estimation, as with ratio estimation, we get two results, the estimated total (and mean) and the regression coefficient. However, the latter is not the same as a ratio. The regression coefficient measures the change in the primary variable that results from a unit change in the auxiliary variable; the ratio measures the proportional relationship between the sum of the primary variable and the sum of the auxiliary variable.

In our example, we would interpret the regression coefficient as follows. For every $100 increase (or decrease) in the agency purchase cost, the agency's amount should have increased (or decreased) by $85.91. If the proportional relationship between the primary variable and the auxiliary variable is needed, the ratio estimate must be used. Using this method, we could also say that we estimate that the total cost that the agency should have spent on small purchases was $4,374,276 and we are 95-percent confident that the true but unknown total is $4,374,276 plus or minus $68,911.02.

Difference Estimation

From our sample of 50 small purchases, we would estimate that the total that the agency should have paid was $4,394,000, and we are 95-percent confident that if we had examined all 10,000 small claims, then the total value should range between $4,394,000 - $136,816, or $4,257,184, and $4,394,000 + $136,816, or $4,530,816.

Which Method to Use

Using the information from the computer printout, the appropriate technical staff will help make a decision on which method of estimation best fits the data and the objective of the job. The evaluator can use the SRO-STATS program and test various sample sizes but may want to calculate the necessary sample size; the formula is a modification of the formula given earlier, but the evaluator should go to either the technical staff or one of the advanced textbooks listed in the bibliography.

Other Advantages and Disadvantages of Ratio, Regression, and Difference Estimation

One big advantage of ratio, regression, and difference estimation is that they adjust the sample results to known population data when we compute totals. If the sample mean for the auxiliary variable turns out to be lower than the population mean for the same variable, the sample results are adjusted upward. If, however, the sample mean for the auxiliary variable turns out to be higher than the population mean, sample results are adjusted downward.

The first major problem in using the three advanced methods of estimation is that the evaluator who does not capture the auxiliary variable total at the time of sample selection and who goes to the agency to get that information may find that the agency has updated its computer records or files so that the total needed to calculate these methods is not available.

The sample size should be large enough to include at least 10 nonzero differences; some statisticians suggest at least 30 nonzero differences. The nonzero differences are necessary so that an estimate of the variability of the data can be made. If the differences are not very frequent, the sample size necessary to get the nonzero differences may be too large to be feasible. However, if no differences are found, the evaluator should report that fact directly rather than increasing the sample size so as to produce an acceptable estimate.

The sample must include a certain minimum percentage of nonzero differences—in general, 10 percent for sample sizes up to 300; for sample sizes greater than 300, the number of nonzero differences should be equal to 30 plus 5 percent of the sample size that is above 300.

The biggest restriction in using these methods is that using the agency's book value total in computing the estimates means that the evaluator should do a separate audit check to verify the total book value figure.

Sampling in the Audit Environment

This chapter is directed more toward financial and management auditing than toward program evaluation. However, some of the points here also apply to program evaluation. In this chapter, we briefly discuss discovery and acceptance sampling, the relationship between audit judgment and statistical sampling, and the characteristics of a good sample.

Discovery Sampling

Discovery sampling is a type of sampling that has a specified probability of including at least one item that occurs very rarely in the population. It is used when there is a possibility of finding such things as fraud and avoidance of internal controls. In discovery sampling, the evaluator can specify the probability of including at least one item with a particular characteristic, if the characteristic occurs at a specified rate in the population. If the sample does not turn up an item with this characteristic, the evaluator can make a probability statement that the characteristic's rate of occurrence is less than that specified.

Discovery sampling can be regarded as a special case of attribute sampling. However, in its usual applications, it does not yield an estimated rate of occurrence, and usually it is used only if the particular characteristic's rate of occurrence is thought to be very small—that is, close to zero. For example, discovery sampling is usually used in financial audits to guard against an intolerable rate of fraud.

The evaluator must specify two things: the rate of error, fraud, or abuse that would be intolerable and the probability of finding at least one occurrence in the sample (if the rate of occurrence is even this high). The required sample size can usually be looked up in published tables like tables 6.1 and 6.2.

For example, assume that the population size is 500 and evaluators want to be 95-percent confident that, if the error rate in the population is 1 percent, they will find at least 1 error in the sample. If the specified intolerable error rate is 1 percent, they would calculate the total errors in the population as 1 percent of 500, or 5. In table 6.1, reading down the column for 5 errors to the row that corresponds to a probability of 95 percent, we find that a sample size of 225 is required.

Then the evaluators select a simple random sample of items and examine each item until they find one with an error or until they have examined the entire sample and found none. If they find an error, they know that the error rate is at least as great as the specified intolerable rate and can extend the review perhaps to the entire population. If they find no deficiencies, they can conclude that the rate of occurrence is less than that specified as intolerable.

An advantage of discovery sampling is that the probability of finding at least one error will increase if the rate of occurrence is greater that the intolerable rate specified by the evaluators. Thus, the likelihood of more quickly finding the one error in the sample is increased, and the average sample size that actually has to be examined is smaller.

Table 6.1: The Probability of Finding at Least One Error in a Population of 500 for Various Numbers of Errors and Sample Sizes[a]

Sample	Total errors in universe size of 500							
size	1	2	3	4	5	10	15	20
5	1.0	2.0	3.0	4.0	4.9	9.6	14.2	18.5
10	2.0	4.0	5.9	7.8	9.6	18.4	26.5	33.8
15	3.0	5.9	8.7	11.5	14.2	26.5	37.1	46.3
20	4.0	7.8	11.5	15.1	18.5	33.8	46.3	56.5
25	5.0	9.8	14.3	18.6	22.7	40.4	54.2	64.9
30	6.0	11.7	17.0	22.0	26.7	46.5	61.0	71.7
35	7.0	13.5	19.6	25.3	30.5	51.9	66.9	77.3
40	8.0	15.4	22.2	28.4	34.2	56.9	71.9	81.8
45	9.0	17.2	24.7	31.5	37.7	61.4	76.2	85.4
50	10.0	19.0	27.1	34.5	41.1	65.5	79.9	88.4
55	11.0	20.8	29.6	37.4	44.3	69.2	83.0	90.7
60	12.0	22.6	31.9	40.1	47.4	72.5	85.7	92.6
65	13.0	24.3	34.2	42.8	50.3	75.5	88.0	94.2
70	14.0	26.1	36.5	45.4	53.1	78.2	89.9	95.4
75	15.0	27.8	38.7	47.9	55.8	80.6	91.6	96.4
80	16.0	29.5	40.8	50.3	58.3	82.8	93.0	97.2
85	17.0	31.1	42.9	52.7	60.0	84.8	94.2	97.8
90	18.0	32.8	44.9	54.9	63.1	86.5	95.1	98.3
95	19.0	34.4	46.9	57.1	65.3	88.1	96.0	98.7
100	20.0	36.0	48.9	59.2	67.4	89.5	96.7	99.0
125	25.0	43.8	57.9	68.5	76.4	94.5	98.8	99.7
150	30.0	51.0	65.8	76.1	83.3	97.3	99.6	99.9
175	35.0	57.8	72.6	82.3	88.5	98.7	99.9	100.0
200	40.0	64.0	78.5	87.1	92.3	99.4	100.0	
225	45.0	69.8	83.4	90.9	95.0	99.8	100.0	
250	50.0	75.1	87.6	93.8	96.9	99.9	100.0	
275	55.0	79.8	91.0	96.0	98.2	100.0		
300	60.0	84.0	93.7	97.5	99.0	100.0		
325	65.0	87.8	95.8	98.5	99.5	100.0		
350	70.0	91.0	97.2	99.2	99.8	100.0		
375	75.0	93.8	98.5	99.6	99.9	100.0		
400	80.0	96.0	99.2	99.8	100.0			
425	85.0	97.8	99.7	100.0				
450	90.0	99.0	99.9	100.0				
475	95.0	99.8	100.0					
500	100.0	100.0						

Total errors in universe size of 500								
25	30	40	50	75	100	200	300	500
22.7	26.7	34.2	41.1	55.8	67.4	92.3	99.0	100.0
40.4	46.5	56.9	65.5	80.6	89.5	99.4	100.0	
54.2	61.0	71.9	79.9	91.6	96.7	100.0		
64.9	71.7	81.8	88.4	96.4	99.0	100.0		
73.2	79.5	88.2	93.3	98.5	99.7	100.0		
79.5	85.3	92.4	96.2	99.4	99.9	100.0		
84.4	89.4	95.2	97.8	99.7	100.0			
88.2	92.4	96.9	98.8	99.9	100.0			
91.1	94.6	98.0	99.3	100.0				
93.3	96.2	98.8	99.6	100.0				
95.0	97.3	99.2	99.8	100.0				
96.2	98.1	99.5	99.9	100.0				
97.2	98.7	99.7	99.9	100.0				
97.9	99.1	99.8	100.0					
98.5	99.4	99.9	100.0					
98.9	99.6	99.9	100.0					
99.2	99.7	100.0						
99.4	99.8	100.0						
99.6	99.9	100.0						
99.7	99.9	100.0						
99.9	100.0							
100.0								

[a]Probability is 100.0

Source: U.S. Air Force, Auditor General, Handbook of Practical Sampling Procedures for Internal Auditors (Norton Air Force Base, Calif.: 1966), p. 125

Table 6.2: The Probability of Finding at Least One Error in a Population of 600 for Various Numbers of Errors and Sample Sizes

Sample	Total errors in universe size of 600							
size	1	2	3	4	5	10	15	20
5	0.8	1.7	2.5	3.3	4.1	8.1	11.9	15.6
10	1.7	3.3	4.9	6.5	8.1	15.6	22.5	28.9
15	2.5	4.9	7.3	9.7	11.9	22.5	31.9	40.2
20	3.3	6.6	9.7	12.7	15.6	28.9	40.2	49.8
25	4.2	8.2	12.0	15.7	19.2	34.9	47.6	57.9
30	5.0	9.8	14.3	18.6	22.7	40.4	54.1	64.8
35	5.8	11.3	16.5	21.4	26.0	45.4	59.9	70.5
40	6.7	12.9	18.7	24.2	29.3	50.1	64.9	75.4
45	7.5	14.4	20.9	26.9	32.4	54.4	69.4	79.5
50	8.3	16.0	23.0	29.5	35.4	58.4	73.3	83.0
55	9.2	17.5	25.1	32.0	38.3	62.1	76.8	85.9
60	10.0	19.0	27.1	34.5	41.1	65.4	79.8	88.3
65	10.8	20.5	29.1	36.9	43.7	68.5	82.5	90.3
70	11.7	22.0	31.1	39.2	46.3	71.4	84.8	92.0
75	12.5	23.5	33.1	41.5	48.8	74.0	86.8	93.4
80	13.3	24.9	35.0	43.7	51.2	76.4	88.6	94.6
85	14.2	26.3	36.8	45.8	53.5	78.6	90.2	95.5
90	15.0	27.8	38.6	47.9	55.8	80.6	91.5	96.3
95	15.8	29.2	40.4	49.9	57.9	82.4	92.7	97.0
100	16.7	30.6	42.2	51.9	59.9	84.1	93.7	97.6
125	20.8	37.4	50.4	60.8	69.0	90.5	97.1	99.1
150	25.0	43.8	57.9	68.5	76.4	94.5	98.7	99.7
175	29.2	49.9	64.5	74.9	82.3	96.9	99.5	99.9
200	33.3	55.6	70.4	80.3	86.9	98.3	99.8	100.0
225	37.5	61.0	75.7	84.8	90.6	99.1	99.9	100.0
250	41.7	66.0	80.2	88.5	93.3	99.6	100.0	
275	45.8	70.7	84.2	91.5	95.4	99.8	100.0	
300	50.0	75.0	87.6	93.8	96.9	99.9	100.0	
325	54.2	79.0	90.4	96.5	98.0	100.0		
350	58.3	82.7	92.8	97.0	98.8	100.0		
375	62.5	86.0	94.8	98.1	99.3	100.0		
400	66.7	88.9	96.3	98.8	99.6	100.0		
425	70.8	91.5	97.5	99.3	99.8	100.0		
450	75.0	93.8	98.5	99.6	99.9	100.0		
475	79.2	95.7	99.1	99.8	100.0			
500	83.3	97.2	99.5	99.9	100.0			
550	91.7	99.3	99.9	100.0				
600	100.0	100.0	100.0					

Total errors in universe size of 600								
25	30	40	50	75	100	200	300	500
19.2	22.7	29.3	35.4	48.8	59.9	86.9	96.9	100.0
34.9	40.4	50.1	58.4	74.0	84.1	98.3	99.9	100.0
47.6	54.1	64.9	73.3	86.8	93.7	99.8	100.0	
57.9	64.8	75.4	83.0	93.4	97.6	100.0		
66.3	73.0	82.8	89.2	96.7	99.1	100.0		
73.0	79.4	88.0	93.1	98.4	99.6	100.0		
78.4	84.3	91.7	95.7	99.2	99.9	100.0		
82.8	88.0	94.3	97.3	99.6	99.9	100.0		
86.3	90.9	96.0	98.3	99.8	100.0			
89.2	93.1	97.3	98.9	99.9	100.0			
91.4	94.8	98.1	99.3	100.0				
93.2	96.1	98.7	99.6	100.0				
94.7	97.1	99.1	99.8	100.0				
95.8	97.8	99.4	99.8	100.0				
96.7	98.4	99.6	99.9	100.0				
97.4	98.8	99.7	99.9	100.0				
98.0	99.1	99.8	100.0					
98.4	99.3	99.9	100.0					
98.8	99.5	99.9	100.0					
99.1	99.6	99.9	100.0					
99.7	99.9	100.0						
99.9	100.0							
100.0								

[a]Probability is 100.0.

Source: U.S. Air Force, Auditor General, Handbook of Practical Sampling Procedures for Internal Auditors (Norton Air Force Base, Calif.: 1966), p. 126.

Acceptance Sampling

Acceptance sampling provides us with the decision of whether to accept or reject a specific population. It also assures us that "on the average" very bad populations will be rejected and very good populations will be accepted. In acceptance sampling, a random sample of items is drawn from a population (or "lot"), and the sample is examined or tested. On the basis of this examination, the decision is made to accept or reject the entire lot. The decision may also be made to draw one or more additional samples if the results of the first sample are inconclusive. These are called double, or multiple, acceptance sampling plans, terminology that is derived from the field of industrial quality control. If the number of deficiencies found in the sample is greater than a predetermined number, the entire lot is rejected. If the number of deficiencies in the sample is equal to or less than the predetermined number, the entire lot is accepted.

Acceptance sampling is a variation of attribute sampling, but it does not permit the estimation of the rate of occurrence of the deficiencies.

To use acceptance sampling, the evaluators must specify four criteria:

1. the limit of acceptable quality, or the maximum percentage of deficiencies that can be considered satisfactory on the average over the long run. It should have a high probability of acceptance or a low probability of rejection;

2. the lot tolerance percent defective (LTPD), or the maximum percentage of defects that can be tolerated in a lot. It should have a low probability of acceptance or a high probability of rejection;

3. the probability of incorrectly rejecting a lot of acceptable quality (sometimes called producer's risk);

4. the probability of incorrectly accepting a lot of unacceptable quality (sometimes called consumer's risk).

Once evaluators have specified these criteria, they consult tables of acceptance sampling plans for the plan that comes closest to the criteria. (For an example, see table 6.3 on the next page.) The plan will give the sample size required and the maximum number of acceptable defectives in the sample, referred to as the acceptance number.

To give an example of an acceptance sampling plan, let us assume that the evaluators want to verify the accuracy of the keypunching of the results of questionnaires. For this example, let us assume that the number of questionnaires keypunched is 2,500. The evaluators have decided to define an error as any case containing one or more errors. The evaluators want to take no more than a 5-percent risk of incorrectly rejecting a set of data whose error rate is 1.75 percent, and they want to take only a 10-percent risk of incorrectly accepting a data set whose error rate is 5.8 percent. Therefore, the lot size is 2,500, the acceptable quality limit is 1.75 percent, the LTPD is 5.8 percent, the producer's risk is 5 percent, and the consumer's risk is 10 percent.

As shown in table 6.3, the sampling plan for lot size 2,001 to 3,000, and an acceptable quality limit of 1.61 to 2 percent will meet these criteria. The evaluators will take a random sample of 180 records from the data set and verify the keypunched data against the questionnaire, and they will accept the lot if the number of records in error is 6 or less; otherwise, the lot will be rejected.

Table 6.3: An Acceptance Sampling Table[a]

	Acceptable quality limit					
	1.21-1.60 percent			1.61-2.00 percent		
Lot size	Sample size	Acceptance number[b]	LTPD	Sample size	Acceptance number[b]	LTPD
1-15	All	0		All	0	
16-50	14	0	13.6	14	0	13.6
51-100	16	0	12.4	16	0	12.4
101-200	35	1	10.5	35	1	10.5
201-300	37	1	10.2	37	1	10.2
301-400	38	1	10.0	60	2	8.5
401-500	60	2	8.6	60	2	8.6
501-600	60	2	8.6	60	2	8.6
601-800	65	2	8.0	85	3	7.5
801-1,000	65	2	8.1	90	3	7.4
1,001-2,000	95	3	7.0	120	4	6.5
2,001-3,000	120	4	6.5	180	6	5.8
3,001-4,000	155	5	6.0	210	7	5.5
4,001-5,000	155	5	6.0	245	8	5.3
5,001-7,000	185	6	5.6	280	9	5.1
7,001-10,000	220	7	5.4	350	11	4.8
10,001-20,000	290	9	4.9	460	14	4.4
20,001-50,000	395	12	4.5	720	21	3.9
50,001-100,000	505	15	4.2	955	27	3.7

[a]For this table, according to Dodge and Romig, the risk of incorrectly rejecting a lot whose average quality limit is indicated by the column headings is about 5 percent. The risk of incorrectly accepting a lot whose percentage of defectives equals an entry in the LTPD columns is, at most, 10 percent.

[b]According to Dodge and Romig, accept the lot if the number of defectives does not exceed this value.

Source: H. F. Dodge and H. G. Romig, Sampling Inspection Tables, Single and Double Sampling, 2nd ed. (New York: John Wiley and Sons, 1959).

At first glance, acceptance sampling seems attractive because tabulated plans are available and the sample sizes are usually smaller than those required for estimation sampling. However, the assumptions underlying acceptance sampling make it unusable for some GAO work. The major assumption is that many samples are drawn for a continuous stream of homogeneous items (such as ball bearings or artillery shells) being produced or processed by a system under some form of control. Thus, the decision whether to accept or reject many lots over the long run is generally in accordance with the four criteria described above, but this may not be so with a single lot. The policy or oversight researcher's opinion must be based on the results of a single test; the lot being examined may have been produced by several different processes or departments; and controls, if any, may vary from department to department.

Also, the industrial sampler has little concern about moderately bad situations, but they are important to evaluators because they may indicate fraud or collusion. Further, evaluators, unlike industrial samplers, cannot merely send back a "bad" lot to be reworked at no cost. If they reject a lot because of a moderately bad situation, they may require an unnecessary extension of the test that may cost GAO, or the agency, money. The evaluators should take all these factors into account before deciding to use acceptance sampling.

Acceptance sampling formulas and tables do not permit the development of statistical estimates. However, once an entire acceptance sample has been evaluated, it can be used to develop statistical estimates, if it is a large enough sample.

Audit Judgment and Statistical Sampling

The charge has been made that statistical sampling prevents evaluators from using professional judgment in conducting reviews. This is not correct; statistical sampling is merely a tool to help them make wise decisions. The evaluators still decide what type of review to make, how and when to use sampling, and how to interpret the results. In applying statistical sampling techniques to audit testing, evaluators must make the following decisions that involve professional judgment.

1. They must define the problem. They must decide what to measure, what type of information will provide sufficient facts for the formation of an opinion, and what testing procedures to use.

2. The level of confidence must be specified. This is the probability that an estimate made from the sample will fall within a stated interval of the true but unknown value for the population as a whole. Auditors or evaluators may think of it as the percentage of time that a correct value (within the specified precision limits) will result from using an estimate based on a sample.

3. They must define the population for size and other characteristics. They decide what type of items will be included and excluded, and they specify the time period to be covered.

4. The areas susceptible for sampling must be determined. They should comprise numerous items or similar transactions that can be measured. The evaluators' assessment of the internal control system for an area may determine whether statistical sampling is appropriate. A strong internal control system, for example, may reduce testing to the minimum necessary for verification and may, therefore, call for a different sampling plan or no statistical sampling at all. Prior experience, as well as information from prior evaluations, plays a role here.

Prior evaluations may suggest that certain kinds of records are more prone to error and need higher verification rates than other kinds of records. Thus, evaluators may have to stratify the population between records likely to have a high error rate and those likely to have a low error rate.

5. The maximum error rate that the evaluators will consider acceptable must be decided, as well as the definition of an error. Or, if the evaluators are attempting to estimate the value of balance sheet amount, they must determine the required precision of the estimate in terms of the materiality of the amount being examined and the overall objective.

6. Conclusions about the population must be drawn from the sampling results. In arriving at these conclusions, the evaluators must judge the significance of the errors they have discovered.

Because statistical sampling provides more and better information, it permits greater use of professional judgment and enables evaluators to more effectively analyze the results of tests. And by reducing the work load, statistical sampling allows more time to use professional judgment.

The Characteristics of a Good Sample

In traditional estimation sampling, the ideal sample is characterized as representative. That is, the sample produces an unbiased estimate of the true population characteristic, and this estimate is as precise as possible given the resources available for designing the sample and collecting the data.[1]

[1] This section is based on Ijiri and Kaplan (1970), pp. 42-44. Their article is highly recommended for everyone doing sample design for financial audits and for economy, efficiency, and effectiveness audits.

When we sample for audit purposes, we should expect a good sample to be not only representative but also corrective, protective, and preventive.

As noted above, "representative" means that the sample estimates the true population characteristic as accurately as possible. For example, if we were taking a sample from an inventory of 2,000 items whose true (but unknown) total value was $860,000, we would like the estimated total computed from our sample data to be as close to $860,000 as possible. If the same inventory had a true (but unknown) error rate of 5 percent, we would like our sample to estimate an error rate as close to 5 percent as possible. From our previous discussion, we know that if the population is defined correctly, if the list of population items from which we draw the sample is correct, and if we use random procedures to select the sample items, the sample will be representative. The estimate will be unbiased and have a measurable precision and confidence level.

"Corrective" means that the sample will locate as many error items as possible, so that they can be corrected. Even if the system that generated the errors is not corrected, as many specific instances as possible will be. (Regarding this characteristic, some may state that it is not GAO's job to do the agency's work. However, if we suspect a problem and if, by careful sample design, we can disclose a large number of problem items, our findings may be more effective and more likely to bring about corrective action.)

In the example above, we assumed that the true error rate in the population was 5 percent. If we selected a random sample of 100 items from the population, we would expect to find only 5 error items. However, if we could identify in advance the error-prone items—that is, the items most likely to contain errors based on what they were, how they were stored, how they were accounted for, or the like—we could perhaps isolate them from the other items and take all our sample or the largest part of our sample from these items. Thus, we could maximize the number of errors disclosed by the sample.

"Protective" means that the person who does the sampling attempts to include the maximum number of high-value items in the sample. This approach is common in auditing when auditors isolate the high-value items from the rest of the population, gather data on all these items, and gather data from a sample of the remaining items. Continuing with our inventory example, if we knew that 100 of the items had values in excess of $1,000, we might audit all these items and audit a sample of the remaining items. Or if, in addition to knowing that 100 of the items had values in excess of $1,000, we knew that 500 items had values of $100 to $1,000, we might review half of the 500 items and all the items that had values in excess of $1,000.

"Preventive" means that the sampling method gives agency managers no idea which items will be selected during a review.

When designing a sampling plan, evaluators should keep in mind the desirability of obtaining a sample that is as representative, corrective, protective, and preventive as possible. To do this, they should stratify the population on the basis of dollar value and the likelihood that the items contain errors, use some random method to select the sample from each stratum, and weight the results from each stratum to compute overall estimates for the population.

It is not possible, however, to optimize all four characteristics in a single sample. Instead, a balance must be struck, depending on which characteristic is most important in view of the job objective. Also, in certain types of jobs, one or more characteristics may not require consideration at all.

Random Selection Procedures

One of the most critical parts of every sampling operation is the actual random selection of the units to be examined. Mistakes in estimating the sample size or in evaluating the sample results can be corrected or appropriate adjustments can be made before, or sometimes even after, data collection has been completed. However, a mistake in the sample selection process can materially distort or even invalidate the sample results, particularly if the mistake is not detected, and it can sometimes make it necessary to redo or abandon the work. This chapter describes the steps involved in sample selection. (A more detailed description of selection procedures, the various problems that may confront the sampler, and the methods of overcoming these problems is in appendix II.)

As we have noted, developing a sampling plan is iterative. However, we must assume that certain components of the plan are fixed, even though they might change in a later iteration. Therefore, we assume that the

- audit or evaluation questions have been formulated,
- audit or evaluation strategy has been chosen,
- population of interest has been defined,
- sample design has been chosen,
- sampling units have been defined,
- sample size has been determined, and
- data collection and analysis plans have been made.

In order to make intelligent decisions about the last five of these points, we must have rather detailed knowledge about the physical location and accessibility of the population and a good estimate of the number of items in the population, if the exact number is not available. We must know the practical aspects of gathering the data. We must know whether the sampling units are documents, school pupils, spare parts, or the like; whether the population is at one location or at several locations some distance apart; and whether the sampling units are in file drawers, storage bins, or neighborhoods. And we must know how the measurements will be administered. To individuals? To groups? In person? By telephone? By mail?

The Selection of Sampling Units

There are three basic procedures for selecting statistical samples: systematic selection with a random start, selection based on randomly selected combinations of the terminal digits of the randomly determined portion of an assigned identification number, and random number sampling by computer or by hand.

Systematic Selection With a Random Start

Because of its simplicity and usefulness in many situations, systematic selection with a random start will be discussed first. In this selection procedure, the sample is selected from the population on the basis of a fixed, or uniform, interval between the sampling units, after a random starting point has been determined. The uniform interval between units is obtained by dividing a given sample size into the population size and dropping any decimals in the result. The random start is selected by any method of selecting random numbers (usually a computer program) and is a random number that is between 1 and the uniform sampling interval inclusive. The random start ensures that, for all practical purposes, all sampling units in the population have an equal opportunity of being selected.

For example, assume that we want to draw a sample of 200 items from a file containing 10,100 items. Dividing the sample size into the population size gives a quotient of 50.5. Rounding downward gives a sampling interval of 50. Using the GAO random number program, we would select a number between 1 and 50. Suppose the random starting point between

1 and 50 is 36. We start with item number 36 and pull every 50th item thereafter; the 36th item, 86th item, 136th item, and so on will constitute the sample.

Situations Calling for Systematic Sampling

Systematic sampling with a random start may be used when the sampling units are not numbered or when it would be too cumbersome to attempt to match the sampling units against random numbers. Here are some circumstances in which systematic selection may be used advantageously:

1. The sampling units are long lists or pages of lists.

2. The sampling units are files on records that are not serially numbered, or if they are numbered, they are not in numerical sequence.

3. The sampling units are not suitably numbered and are intermingled with other items that are not to be included in the sample.

4. The sampling units are numbered in blocks of numbers, and some blocks are not used.

Cautions in Using Systematic Selection

To obtain a sample size that is neither too large nor too small, the evaluator must know or be able to closely estimate the population size. Before using systematic selection, the evaluator should make sure that the sample will be drawn from the entire population. If the sample is to be drawn from a list of items, the list must be complete; if the sample is to be drawn from a file cabinet, all the folders must be in the cabinet, or charge-out cards or some similar system must be used to mark the position of missing folders. Otherwise, the evaluator must make special arrangements to ensure that missing sampling units have the opportunity of being selected.

In certain types of populations, the items are arranged so that certain significant characteristics recur at regular intervals. This is called periodicity. Some examples are daily highway traffic passing through a certain intersection during the day and department store sales during a week. A systematic sample with a random start might consist of a certain time of day from the former type of population or a certain day of the week from the latter, even though every time point and every day would have an equal opportunity of being selected. Obviously, samples like these would be unrepresentative of the entire population.

Another example is a population consisting of a payroll list on which every 25th employee is a supervisor. A systematic sample of every 25th name or multiple thereof could result in a sample consisting entirely of supervisors or, more likely, in a sample that excluded all supervisors. Obviously, neither situation is desirable. Discussions with agency personnel commonly disclose situations of this type.

In general, this problem can be minimized if the sample is taken from lists of persons arranged alphabetically by name or in order of Social Security number or from lists of inventory items in sequence by stock number or by the dates the items were first stocked. Before using systematic selection, it is imperative that the evaluator determine whether there is a relationship between the arrangement of the population and the characteristic being measured.

When systematic selection with a random start is used, the selection process must be continued throughout the entire population, as originally defined by the evaluator, even though the population size may have been underestimated when the sampling interval was calculated and even though the selection will produce a larger sample than required. Under no circumstance should evaluators stop when they reach the required sample size. This is equivalent to "throwing out" part of the population and could result in an unrepresentative sample. If the sample turns out to be too large, it can be reduced by using one of the procedures described in appendix II.

Once the sample has been selected, it is not permissible to substitute other items for sample items that are missing (that are out of the file or the like) or for sample items that may not have adequate supporting material to permit measurement. Every effort should be made to locate the missing items or supporting material. If they cannot be located, this fact should be noted and reported as one of the sample results.

The evaluator should also be aware that, if the sampling units are arranged in ascending or descending order of magnitude, a systematic sample will yield a smaller measure of precision than a random number sample.

Selection Based on Randomly Selected Combinations of Terminal Digits in Identification Numbers

This is really another method of systematic selection with a random start, but the mechanical procedures for selecting the sample are different. Certain types of sampling units have been assigned consecutive identification numbers. Examples are Social Security numbers, inventory stock numbers, and transaction numbers assigned in the order in which documents were received or processed. The important feature of the identification numbers is that the terminal digits (usually the last three, sometimes the last four) can usually be assumed to be random with respect to the characteristics the evaluator wants to measure.

A sample can be selected from a population of units having such identification numbers by selecting all the items (or persons) having identification numbers ending in a certain randomly selected digit or combination of digits. Because there are 10 digits from 0 through 9, each digit will appear in the last position in approximately 10-percent of the identification numbers. Thus, all identification numbers having a terminal digit that matches a randomly selected digit from 0 through 9 will constitute a random 10-percent sample. Similarly, there are 100 possible combinations of pairs of digits between 00 and 99. Each pair of digits will appear in the last two positions of 1 percent of the identification numbers. Thus, all identification numbers whose last two digits match a randomly selected pair of digits between 00 and 99 will constitute a random 1-percent sample.

The steps in this selection procedure are

1. determining the required sample size,

2. dividing the sample size by the population size to obtain the sampling rate (or percentage), and

3. selecting the required quantity of random digits or combinations of random digits by using some suitable source of random numbers.

The percentage indicates the number of digits that should be in the randomly selected combination of digits and the number of combinations that should be selected. For a 20-percent sample, we would match against 2 randomly selected digits between 0 and 9; for a 30-percent sample, against 3 randomly selected digits between 0 and 9; for a 1-percent sample, against a pair of randomly selected digits between 00 and 99; and for a 3-percent sample, against 3 pairs of randomly selected digits between 00 and 99.

For example, to measure the accuracy of payroll records at an installation employing 6,000 persons, evaluators determine that a sample of 240 records will be adequate. They decide to draw the sample by selecting the payroll records of employees whose Social Security numbers end in certain randomly selected pairs of digits. Because a sample of 240 from a population of 6,000 is a 4-percent sample, the evaluators will need four pairs of digits. Using the GAO random number program, they select the following pairs of digits between 00 and 99: 01, 26, 85, and 94. Then they examine the payroll records of all employees whose Social Security numbers end in those digits.

In this type of sampling, selection should not be based on the leading digits in the identification number, because these digits frequently are codes and are not assigned in serial order.

This type of sampling is sometimes called digital selection, digital sampling, or junior digit sampling.

Random Number Sampling

For a simplified example of random number selection of sampling units, suppose that we want to make random selection of 1 person from a population of 10 people. The evaluator knows that the likelihood of selecting any specific person is 1 in 10. The probability is usually expressed as a proportion, 0.10, or as a percentage, 10 percent. The probability is known because the only factor involved in random selection is chance. Subjective considerations (conscious or otherwise), such as selecting new-looking pay records, people who look approachable, or military installations in nearby locations, are completely avoided.

If the population is very small, such as the 10 people, the sample could be selected by recording some identification symbol on 10 slips of paper (this use of a symbol is so that the slips of paper are the same size). The slips of paper could then be placed in a container and mixed thoroughly, and a blindfolded person could withdraw a quantity of slips equal to the specified sample size. The identification numbers on the slips would indicate which people to select.

Although this selection method is practical only when the population is very small, most random selection methods are merely extensions of it.

For most jobs, the procedures are to (1) have a set of random numbers generated by a computer or (2) use the computer to select randomly from records in machine-readable form.

Programmed random number generators are available for use on most computer systems. This includes the GAO random number program. These generators are designed to produce a selection of random numbers that will be suitable for, or can be adapted to, most numbering systems, including compound numbering systems. For most sampling applications, such generators reduce to minutes the time required for random number selection.

In its simplest form, random number sampling is a selection procedure in which a quantity of random numbers equal to the specified sample size is first selected from either a random number program or a table of random digits and then matched against the serial numbers, stock numbers, transaction numbers, or whatever, that are assigned to the sampling units in the population. If the sampling units are not numbered, the evaluator may develop a numbering system for identifying each unit. For example, if documents are entered in a computer list with 25 lines to the page, documents 1 through 25 could be assigned to the first page, 26 through 50 to the second page, and so on. If the items have their own numbers, the sampling process will be greatly simplified if they are arranged in numerical sequence or if a list of the cases in numerical sequence is available. The sampling units having numbers that correspond to the selected random numbers constitute the sample.

First, the beginning and ending numbers of the items in the population are determined. Then numbers falling between the beginning and ending numbers equal to the specified sample size are selected by either a computer program or a table of random digits. For example, if we want to select a sample of 200 items from a population of 8,894 items numbered from 265 through 9,158 and we were to use the GAO random number program, we would enter into the computer program that the lowest random number was 265 and the highest random number was 9,158 and that we wanted to generate 200 random numbers. For most applications, numbers that duplicate a number that has already been drawn are discarded and the additional random numbers are selected to achieve the sample size. The GAO random number program does this automatically. Any acceptable method of generating random numbers can be used with any purely numerical numbering system and can be adapted for use with an alphabetical-numerical numbering system. In most situations, the use of random number sampling is not as simple as this. (Appendix II gives detailed descriptions of how to adapt random number selection to compound numbering systems and other complicated situations. Statisticians can provide guidance on these procedures.)

When preliminary results indicate that the sample is larger than needed, the evaluator may want to decrease the sample size. Basically, this is done by taking a random sample of the random sample. (Details on this procedure are given in appendix II.)

The Application of Selection Procedures

The selection procedures described in this chapter are applicable to simple random sampling, stratified sampling, and cluster sampling. In simple random sampling, we have only a simple population, and only one procedure is used to select the entire sample.

With stratified sampling, the population is divided up into two or more separate subpopulations, or strata. Thus, a different procedure could be used to select the samples in the various subpopulations. Depending on the arrangement of the items in the subpopulations and the numbering systems employed, it might be advisable to use random number sampling in some of the strata and systematic selection with a random start in others.

In the application of these procedures to cluster sampling, one procedure might be used to select the clusters. If it were necessary to select a sample of items within these clusters, a different procedure might be used to do this. For example, systematic selection with a random start could be used to select the clusters and random number sampling could be used to select the sample items within the clusters.

Selection With Probability Proportional to Size

When evaluators apply random selection procedures to cluster sampling, the clusters can be selected with probability proportional to size (PPS) or to a related variable that can be used as a measure of size. This sampling method is based on the assumption that the variable to be measured is highly correlated with some data already known about the cluster, such as number of inhabitants, dollar volume of transactions, or number of students in a school system. If the assumption is correct, this selection method will yield a smaller sampling error than other methods would.

For example, table 7.1 lists claims-paying offices and the number of claims each one paid in 1982. Suppose we want to estimate the dollar value of the claims that were paid. It is reasonable to assume that the dollar value of claims that were paid is approximately proportional to the number of claims that were paid. Assume that the maximum number of offices that can be audited is 20.

Table 7.1: Selection With Probability Proportional to Size

Office	Number of claims	Range of cumulative numbers: Lower	Range of cumulative numbers: Upper	Random numbers
New York	2,936	1	2,936	01038; 02770; 02471; 01174
Hicksville	1,245	2,937	4,181	
Paterson	471	4,182	4,652	
Bronx	2,335	4,653	6,987	
Atlanta	1,775	6,988	8,762	
Pittsburgh	1,254	8,763	10,016	09745; 09094
Tampa	636	10,017	10,652	
Charlestown	174	10,653	10,826	10679
Chicago	2,562	10,827	13,388	12993
Springfield	1,630	13,389	15,018	14922; 13547; 14300
Cincinnati	687	15,019	15,705	15150
South Bend	139	15,706	15,844	
St. Paul	1,818	15,845	17,662	17237
St. Louis	1,114	17,663	18,776	17850
Columbia	148	18,777	18,924	
Detroit	2,159	18,925	21,083	
Cleveland	1,327	21,084	22,410	
Fort Worth	668	22,411	23,078	22638; 22952
Waco	163	23,079	23,241	23223
San Antonio	430	23,242	23,671	
Nashville	625	23,672	24,296	
Chattanooga	202	24,297	24,498	
Jackson	187	24,499	24,685	
Oakland	1,469	24,686	26,154	25972; 25402
Portland	723	26,155	26,877	
Fresno	281	26,878	27,158	
Los Angeles	2,162	27,159	29,320	

(continued)

Office	Number of claims	Range of cumulative numbers		Random numbers
		Lower	Upper	
Van Nuys	361	29,321	29,681	
San Diego	597	29,682	30,278	29841
Honolulu	169	30,279	30,447	

Source: Interstate Commerce Commission, Bureau of Transport Economics and Statistics, Table of 105,000 Random Decimal Digits (Washington, D.C.: 1949), p. 20, col. 7, line 971, through col. 9, line 952.

First, we set up a range of cumulative numbers of claims for each office, as shown in the third and fourth columns of the table. Next, we select 20 random numbers between 1 and the total number of claims paid—that is, between 1 and 30,447. Then we enter each random number on the line for the office whose range of paid claims includes the random number. (For example, random number 09745 is entered on the line for Pittsburgh.) This identifies the sample office. Note that some of the offices are included in the sample more than once; this is characteristic of PPS sampling. Sampling with replacement is used. Thus, when the random numbers are selected, duplicates should not be eliminated.

Each office's probability of selection is proportional to the number of paid claims. Yet each office, from the smallest to the largest, has an opportunity of being selected.

This example shows only how the sample would be selected. The major use of PPS sampling is in two-stage sampling when the cluster sizes vary greatly, as they do here. If clusters were chosen with equal probability, the variation in cluster sizes would increase the computed variation between clusters and thus the overall precision of the estimate. Using two-stage sampling and PPS sampling to select the primary units, the evaluator can calculate subsampling rates within the primary units in a way such that the second-stage sample sizes within each primary unit are equal and, at the same time, the sample is self-weighing. Therefore, the sample can be treated as if it were a simple random sample of cluster, which greatly simplifies the calculations.

The formulas for computing estimates and sampling errors for samples selected with PPS are beyond the scope of this paper.

A Final Check

Once the selection procedure has been decided upon, the procedures to be followed should be written in a sampling plan. Ordinarily, no deviations from the plan should be permitted during the selection process. If unforeseen circumstances make it necessary to modify the sampling plan, the circumstances as well as the modified procedure should be described in the working papers.

Before leaving the field, the evaluators should review the entire sample selection process to ensure that they

1. correctly defined the population,

2. drew the sample from the entire population as originally defined,

3. did not substitute readily accessible sampling units for units that would have been difficult to locate or question, and

4. correctly recorded the pertinent information on each selected unit.

Data Collection and Analysis Considerations Related to Sampling

Sampling is a precursor to data collection. In this chapter, we briefly review some basic ideas about data collection. Data collection methods and the types of data gathered vary with the type of application. Six examples of applications from various disciplines follow.

1. In evaluation and policy analysis, the sampling units are often people who are interviewed, either in person or by telephone, or who are asked to fill out mailed questionnaires.

2. When data are collected from the general public, the housing unit is often the only means of getting at the persons in the sample. When this is so, we have a cluster sampling situation in which the housing unit is the cluster. All the people in the household come into the sample at once.

3. In accounting, data are quite frequently gathered from documents such as vouchers, purchase orders, and ledgers or from computer files. Sometimes the data are gathered by actually counting or measuring and pricing items, as in verifying a physical inventory.

4. To verify various types of book balances, evaluators may have to send letters to the persons listed on an agency's records in order to determine that the agency's information is correct. In verifying inventory held in public warehouses, the evaluators might send confirmation letters to the management at each warehouse, asking them to say how much inventory is held in the warehouse.

5. When agricultural data are collected, quite often measurements are made in the field, such as measuring the procedures used by Department of Agriculture meat inspectors or the number of Chilean grapes that are contaminated. Alternatively, such data as the price farmers received for grain may be

collected from farmers' records or from grain elevator operators.

6. In industry, data are frequently collected by physical measurements, such as testing the tensile strengths of wire or measuring the temperature at which a thermostatic switch will operate.

Checking the Quality of Data Collection

When data are collected, every effort should be made to examine the correct sampling units, to collect and record the data correctly, and to have another person independently verify all measurements and computations. If the data collection operation is large, it is usually advisable to have qualified, well-trained personnel make a quality review of a randomly selected subsample of the sampling units verifying the measurements and computations. This permits a statistical measurement of the data collection errors that can be used to check the accuracy of the final sample results.

If the data are to be collected by questionnaire (including confirmation letters), the evaluators must remember that once they have placed the questionnaire in the mail, for all practical purposes control is lost, since it is impossible to further explain the questions and instructions. What comes back depends on how the respondents react to the questionnaire and how well they understand it. Therefore, the instrument should be as short as possible, it should look easy to answer, and the questions and instructions should be as simple as possible. The questionnaire should be pretested under conditions that are as similar as possible to those under which it will be answered. With rare exceptions, questionnaires should be designed by a team that includes a subject-matter specialist and a questionnaire design specialist. (See transfer paper 10.1.7 entitled Developing and Using Questionnaires, first issued in July 1986.)

If the data are to be collected in person or by telephone, the interviewers should be given a complete set of instructions or an interviewer's manual. They should use an interview form, a questionnaire designed by the team and read to the respondents so that all the interviewers ask the same questions. The interviewers should be trained to read each question exactly as it was written and to provide explanations and clarifications if the respondents do not understand it. Also, interviewers should be trained to maintain a friendly but neutral attitude and not inject personal opinions that might lead the respondents into giving specific answers. (See Using Structured Interviewing Techniques, listed in "Papers in This Series.")

Even with the best training and in the best circumstances, each interviewer has mannerisms that affect people differently and thus affect their responses. This is called interviewer bias. To attempt to overcome it, interviewer assignments should be randomized so that a single interviewer is not responsible for all interviews in a particular area, town, or section of a city. (Travel costs are, of course, a big factor in deciding whether this can be done and how.)

For example, assume that interviews are to be conducted in three small cities A, B, and C, which are fairly close together, and that three interviewers X, Y, and Z are available. To randomize assignments, we randomly assign one third of the interviews in city A to interviewer X, one third to interviewer Y, and one third to interviewer Z, and we follow the same procedure in cities B and C. (More sophisticated schemes for randomizing interviewer assignments and testing interviewer bias are described in the literature.) Randomization of telephone interviewer assignments, of course, is much more easily accomplished.

The Problem of Nonresponse

One of the most troublesome problems in any type of data collection operation—whether it is a sample or a census—is nonresponse, sampling units for which the data cannot be collected. Examples are persons who do not bother to fill in and send back questionnaires and those who send back incomplete questionnaires, persons who are not at home or prefer not to be interviewed in person or by telephone, and documents missing from a file from which the sample is selected. Therefore, we have two types of nonresponse: (1) unit nonresponse in which data for the entire sampling unit is not collected and (2) item nonresponse in which one or more measurements on the sampling units are not made.

It can be a serious mistake to assume that the nonrespondents are like those who do respond. For example, suppose that we want to know whether prisoners at a state penitentiary favor abolition of the death penalty. The population is the 2,500 prisoners at the penitentiary on a certain date. We send each prisoner a questionnaire asking only one question: "Do you favor abolition of the death penalty? (Circle one) yes no." We receive 1,500 completed questionnaires (a 60-percent response rate). A tabulation of the responses is shown in table 8.1.

Table 8.1: Tabulation of Responses to Question on Abolition of the Death Penalty

Response	Number	Percent
Yes	900	60
No	600	40
Total	**1,500**	**100**

From this, we might conclude that the prisoners favor abolition of the death penalty by a majority of three to two. If we then assumed that the nonrespondents looked like the respondents, we might think that 1,500 prisoners were in favor of abolition of the death penalty and 1,000 were not in favor. However, if information were available that permitted us to analyze the type of sentence the respondents had received, we might discover the data in table 8.2.

Table 8.2: Type of Sentence Respondents Received

Response	Death sentence	Other sentence	Total
Yes	600	300	**900**
No	0	600	**600**

Also, let us suppose that the 1,000 nonrespondents were all prisoners who had received another sentence. This additional information might lead us to revise our conclusions about the prisoners' attitudes.

The effect of nonresponses can be overcome by making multiple follow-ups in order to try to reduce the number of nonresponses to, say, 5 percent, unless this percentage includes potential nonrespondents who may have a disproportionate effect on the results (such as several very large firms that fail to respond in an economic survey).

In questionnaire surveys, we can take a random sample (say 25 to 33 percent) of the nonrespondents and attempt to interview them by telephone or in person to obtain answers to at least the key questions. The responses from this sample of nonrespondents then can be weighted so that the results are representative of all nonrespondents. Then, using the stratified approach, we can make an overall estimate using the results from the respondents and the nonrespondents. This resulting overall estimate will be unbiased, but the sampling error will be larger, because responses have not been obtained from the full sample originally selected.

Various statistical tests based on known data can be made for differences between respondents and nonrespondents. Examples of the types of data that can be tested for significant differences are mean or median income, educational level, mean age, and race. If there are no significant differences between the two groups, sometimes it is safe to assume that the nonrespondents look like the respondents, at least with respect to the characteristics we can measure. In personal interview surveys, a more experienced interviewer or supervisor can attempt to interview the persons who were not at home and to convert refusals into responses, or at least try to find out why they refused to participate in the survey. This may enable us to decide whether they have characteristics that are different from those of the typical respondents.

A major mistake is to make arbitrary substitutions for missing sample documents, nonrespondents, and the like. An example of arbitrary substitution is to take the file folder immediately following or preceding a sample folder that is missing from a file; another is to take the household next door to a sample household in which no one was home. The fallacy of this approach is that the record may be missing from the file for a reason, such as fraud or collusion. The household in which no one was home may be occupied by a single person or by a couple who are both employed and have no children. The responses of these people may be different from the responses of those who are usually at home at the time the survey was made. Also, some people may refuse to be interviewed because they are emotionally involved in the survey's subject, and these may be the very people whose answers we want.

Sometimes, random procedures are used to select substitutes for missing or unavailable records or for nonrespondents. Although this is better than arbitrarily selecting the closest available record from a file or the household next door, we are really obtaining more data about the same type of sampling unit—the households that will respond or the records that are available.

Another technique is to assume the missing information was found and had a value for the characteristic we were measuring that was in the agency's favor. With this assumption, we would then make our estimates to the population and see how it affects our conclusions. If the results are sufficient to make our point and support our conclusion, then we can use this assumption to make extremely conservative estimates to the population.

If nonresponse cannot be overcome or if missing records cannot be located, this is a weakness in the evidence and as such must be disclosed in the report. Also, the rate of nonresponse to individual questions should be reported. The user of the information can thus know the actual results and can evaluate them accordingly.

Chapter 9

Statistical Sampling As a Friend of Evaluators

When a design is developed for answering an evaluation question, the question arises as to whether or not to sample. If the population is small or the individual sampling units in the population are very important, it is often advisable to examine every unit in the population. However, if the population is large, a sample is preferred to a census because the information that is wanted can be obtained more cheaply, more quickly, and often more accurately and in more detail. In some instances, only one of these benefits applies, and in some extreme situations, not one does. These points deserve some explanation.

Sampling is usually cheaper than a complete review of the population because, by definition, it usually deals with only a small group selected from the population. The total cost of getting information includes a variable cost related to examining the individual items. By reducing the number of items to be examined, sampling permits a substantial reduction in that variable cost. However, a good sampling plan may add some costs that would not be present in a census. Although almost always much smaller than the saving, such costs should not be ignored. They usually cover

- developing the sampling plan,
- selecting the sample,
- monitoring the sample selection process,
- processing the data and calculating estimates and sampling errors, and
- providing special training or instructions necessary for all but developing the sampling plan.

With regard to speed, sometimes a recommendation must be prepared or a decision must be made within a relatively short time. No matter how good the quality, information is of no help unless it is received in time to be used in making the recommendation or arriving at the decision. The measurement or examination process takes time; so does the summarization of results. Because a sample has fewer items than a census, these processes can be done more quickly in order to make them more useful to the decisionmakers.

Sometimes an attempt will be made to obtain more information from the sample than it was originally designed for. An example of this is taking a sample that was designed to evaluate the effectiveness on a runaway-youth shelter program as a whole and then attempting to develop estimates for different domains of interest (for example, classifying the youths by gender, age, race, or the marital status of their parents). In some cases, attempts to use a sample for purposes other than those for which it was designed can lead to estimates with sampling errors as large as, or larger than, the estimates themselves.

Similarly, many believe that sampling may furnish less accurate answers than a census. We would like to assert the opposite view, even though it may not, at first glance, seem reasonable. The basis for this suggestion is that because of the smaller number of units to be examined, the staff will do a better job of measuring, recording, processing, and reporting. Because sampling involves the observation of fewer units, it frequently allows us to use personnel who have been better trained to collect, process, and evaluate the data than would be practicable in a census of the population. In fact, it has been found that measuring physical inventory by sampling is more accurate than counting and pricing every unit. Also, because fewer observations are needed in sampling, the measuring process can be done more nearly simultaneously and the result is more likely to present correctly the status of the population at a given time than would the result of a census, during which changes may take place. For instance, by the time the last item has been measured, the first one may have been used up or materially changed.

By suggesting that sampling permits more detailed information to be obtained, we mean that if an attempt is made to measure all the units in the population, it may be possible to make only one or a few measurements on each unit. However, if sampling is used and fewer units are selected, it may be possible to collect much more data about each unit and thus develop more information about the population. For example, the Bureau of the Census is mandated to count the population (a complete enumeration) in the decennial census, but the detailed statistics on the demographic characteristics of the population and the nation's housing are developed from a sample of selected households counted during the census.

In sum, if statistical sampling is feasible and is carried out correctly, it usually has advantages over both a census and judgment sampling. When the evaluation objective requires inferences about the population of interest, statistical sampling offers the following advantages:

1. Results from statistical samples are objective and defensible. Mathematical theory provides support and protection for interpreting results when samples are selected by one of the methods in this paper. However, judgment samples cannot be defended by mathematical theory, not because the conclusions that are reached are wrong but because there is no way of objectively determining if they are right or wrong.

2. The precision of a sample result can be expressed in numerical terms.

3. Statistical sampling permits advance determination of sample size on an objective basis.

4. Statistical sampling often permits better planning and budgeting of resources.

5. Results from statistical samples taken at widely dispersed locations, by different evaluators, are easily combined for an overall evaluation.

6. Statistical sampling may save time and money.

7. Using statistical sampling requires an orderly discipline and a greater exercise of professional judgment on the part of the evaluator.

Statistical sampling is not a panacea for all evaluation problems. It has no magic formulas for defining evaluation objectives and the population of interest. It cannot decide which of its many techniques is the most efficient or economical for the problem at hand, and it cannot say which confidence levels and sampling errors are most acceptable. Formulas have no brains and will only do what they are told to do. They provide results but do not interpret what caused them or how useful they might be. Statistical sampling is a precision tool but only a tool. To be used correctly, it requires good professional judgment in every step from defining the evaluation objectives and the population of interest to determining the language and method of presentation of the final report.

The advantages of statistical sampling make it very difficult to challenge its superiority for making inferences about evaluation populations. However, it does not necessarily follow that judgment sampling and census are obsolete and should never be used. Statistical sampling is not needed if evaluators check a few typical units through a system or computer, if they purposely want to examine certain units because they suspect improprieties, or if they are looking for one or two bad examples to illustrate a previously found control weakness. As a general rule, statistical sampling is not needed if the evaluator does not intend to present conclusions about the population.

Statisticians and others rightfully object when judgment sampling is used as a basis for conclusions about the population. They object not because the conclusions may be wrong but because there is no way of knowing or objectively determining if they are right or wrong. Evaluators should understand this limitation and use the technique accordingly.

In the final analysis, the type of sampling used is a matter of judgment. However, the type selected should be consistent with the job objectives, which means it should be defensible for the assignment's purpose.

Theoretical Background

This appendix explains the basis for sample estimates and the concepts of confidence and precision (or sampling error). It is intended for those who have not completed a college course in statistics or who have completed one but need a refresher.

As we noted in chapter 4, we use measurement or observations of a sample to draw inferences about the population from which the sample was drawn. The sample is used to estimate means, totals, and rates of occurrence as well as many other values in the population. Assuming that the sample mean estimates the population mean, we multiply the sample mean by the number of items in the population to estimate the population total. Similarly, when estimating rates of occurrence, we assume that the rate of occurrence found in the sample estimates the rate of occurrence in the population. Such assumptions are based on the laws of probability.

Calculating Probabilities

Let us consider a dice-throwing experiment that uses a pair of perfect dice. For each point that can be thrown, we tabulate the number of ways the point can be made and the probability of making the point in a single throw. (See table I.1.) With this table, a person can calculate the probabilities of all possible outcomes even before picking up the dice. For example, the probability of throwing 2, 3, or 12 (called "craps" or "crapping out") is 11.11 percent (2.78 + 5.55 + 2.78). The probability of throwing 7 or 11 (a "natural") is 22.2 percent (16.67 + 5.55). The probability of throwing a 7 is 16.67 percent. The probability of throwing any one of the points 5 through 9 is 66.7 percent, and the probability of throwing any one of the points 3 through 11 is 94.4 percent.

Table I.1: Dice-Throwing Probabilities

Point	Number of ways point can be made	Probability of making point in a single roll
2	1	2.78%
3	2	5.55
4	3	8.33
5	4	11.11
6	5	13.89
7	6	16.67
8	5	13.89
9	4	11.11
10	3	8.33
11	2	5.55
12	1	2.78
Total	**36**	**99.99%**

The ability to calculate probabilities for a dice-tossing experiment depends on mechanical conditions—the way the dice are constructed and roll. Similarly, it has been found that the probabilities for sampling experiments also depend on certain conditions of sample selection. Tabulated in table I.2 is the outcome of a sampling experiment in which 400 samples, each consisting of 400 beads, were drawn at random from a jar containing 20,000 beads, of which 4,000 (20 percent) were red and 16,000 (80 percent) were blue.

Table I.2: Number and Percent of Red Beads Found in 400 Random Samples of 400 From a Population of 20,000 Containing 4,000 Red Beads

Frequency	Number	Percent
1	55	13.75%
1	58	14.50
2	59	14.75
1	61	15.25
2	62	15.50
2	63	15.75

(continued)

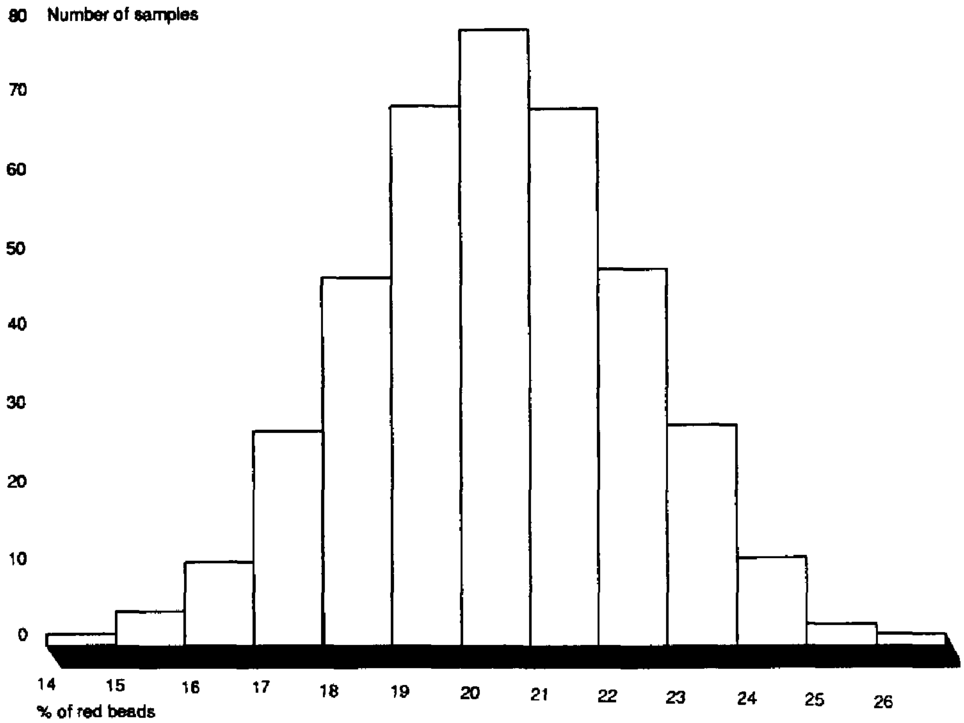
Frequency	Number	Percent
4	64	16.00
2	65	16.25
6	66	16.50
8	67	16.75
7	68	17.00
6	69	17.25
10	70	17.50
14	71	17.75
12	72	18.00
11	73	18.25
19	74	18.50
15	75	18.75
17	76	19.00
18	77	19.25
15	78	19.50
18	79	19.75
23	80	20.00
23	81	20.25
20	82	10.50
16	83	20.75
19	84	21.00
14	85	21.25
9	86	21.50
11	87	21.75
14	88	22.00
15	89	22.25
6	90	22.50
9	91	22.75
7	92	23.00
6	93	23.25
4	94	23.50
2	95	23.75
2	96	24.00

(continued)

Frequency	Number	Percent
4	97	24.25
1	98	24.50
1	100	25.00
2	101	25.25
1	103	25.75

If each sample were an exact image of the population, each would contain 80 red beads (20 percent of 400). Instead, as we can see, the samples vary from as few as 55 red beads (13.75 percent) to as many as 103 red beads (25.75 percent), with the remaining samples having varying quantities of red beads between those two extremes. However, most samples contain quantities of red beads that are close to the number we would expect, knowing what we do about the population. The two categories in which the largest number of samples (23) fell contain 80 and 81 red beads, or 20.00 and 20.25 percent, respectively. Further, 274, or 68.5 percent, of the samples contain between 18 and 22 percent red beads, and 382, or 95.5 percent, of the samples contain between 16 and 24 percent red beads. The results of the samples are shown in figure I.1. As can be seen, the samples are arranged almost symmetrically about the category that contains the true percentage of red beads.

Figure I.1: Number of Samples Classified by Percentage of Red Beads



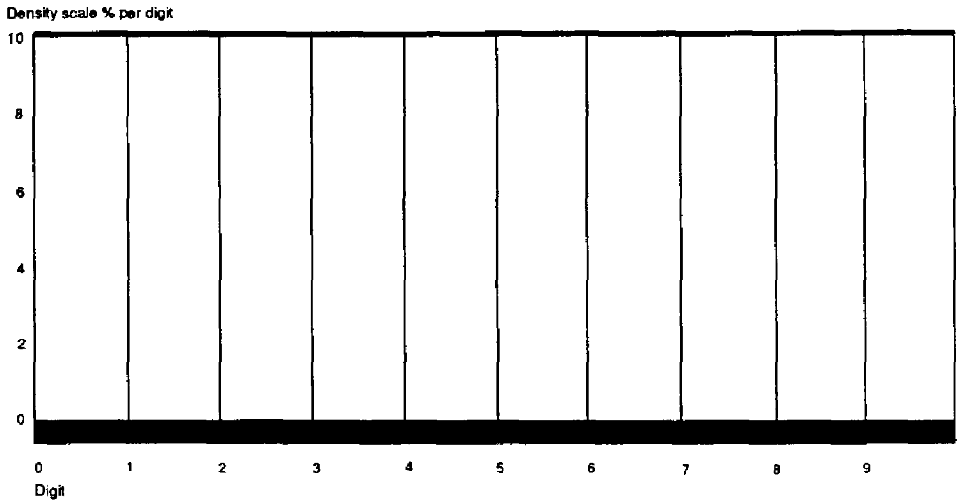
We might view the percentage of red beads in a sample of 400 beads as a shot at knowing the percentage of red beads in the jar, but we would have to understand that the shot is affected by sampling variation. Figure I.1 indicates the confidence that should be associated with each level of precision for this shot. For instance, the shot (the sample percentage of red beads) should be between 18 and 22 percent with about 68-percent confidence. That is, the sample percentage of red beads will "likely" fall within 2 percentage points of the population percentage of red beads. The shot should be between 16 and 24 percentage points with about 95-percent confidence. That is, the sample percentage of red beads will "very likely" fall within 4 percentage points of the population percentage of red beads.

If we look at the red beads as ones and the blue beads as zeros, the percentage of red beads in the population or sample is the mean (in percentage terms) of the ones and zeros. The lesson from figure I.1 is that a sample mean of a large random sample from the jar is a reliable shot at the population mean.

We can illustrate the situation in which we are sampling for variables by considering a very large population of single digits. Each of the 10 digits (0 to 9) constitutes 10 percent of the population. We summarize this population in figure I.2 by setting each digit as a separate class interval.[1] The population mean is 4.5. Note that the arithmetic mean is the point of balance, if the chart is a physical object. A random sample of a single item from this population gives no useful information about the mean. This one item could be any of the 10 digits with an equal chance of occurrence.

[1]We are indebted to Donald T. Gantz of the Department of Mathematical Sciences, George Mason University, Fairfax, Virginia, for developing this example.

Figure I.2: The Relative Frequency of a Large Population of Single Digits Between 0 and 9

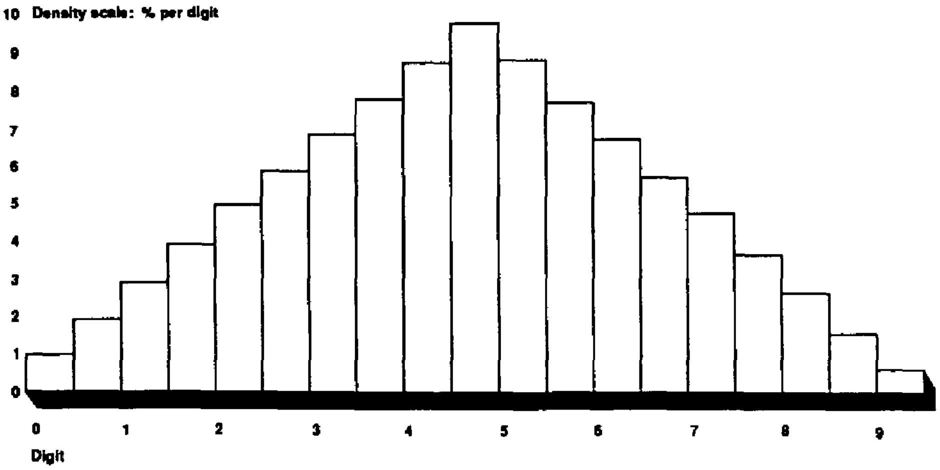


A sample of two elements (n = 2) provides more information about the population mean. To see this, refer to table I.3 and the histogram in figure I.3 for means of samples of two elements from this population. If we use simple random sampling, all the two-item samples have an equal chance of being selected. Of the 19 possible values (0, 0.5, 1, 1.5, . . . , 8.5, 9) for the sample mean from a sample of items, the most likely value is 4.5 (the population mean). The chance that the sample mean will be 4.5 is 10 percent, and the chance that it will be in the range from 4 to 5 is 28 percent. Correspondingly, each of the extreme values 0 and 9 has only a 1-percent chance of being the mean of a sample of two items. Note the strong centralizing effect of averaging only two randomly selected items from the population.

Table I.3: The Relative Frequency of Means of Samples of Two From a Population of Digits Between 0 and 9

Interval	Density
0	1
0.5	2
1.0	3
1.5	4
2.0	5
2.5	6
3.0	7
3.5	8
4.0	9
4.5	10
5.0	9
5.5	8
6.0	7
6.5	6
7.0	5
7.5	4
8.0	3
8.5	2
9.0	1

Figure I.3: The Relative Frequency of Means of Samples of Two From a Population of Digits Between 0 and 9



Now, consider depictions of the means of random samples of three items from the same population, as shown by the frequency data in table I.4 and the histogram in figure I.4. If we use simple random sampling, we find that all the possible three-item samples have the same chance of being selected. Figure I.5 presents the same information as figure I.4 but a smooth curve gives the heights (that is, the density) of the rectangles over the class intervals. Of the 28 possible values (0, 0.333, 0.667, . . . , 8.333, 8.667, 9) for the sample mean of a three-item sample, 4.333 and 4.667 (the values closest to 4.5) are the most likely. There is about a 30-percent chance that the sample mean will be in the range from 4 to 5. Correspondingly, of the extreme values 0 and 9, each has only a 0.1-percent chance of being the mean of a three-item sample. Further, the two extreme values 0 and 1 (as well as the tail values 8 and 9) have only a

3.5-percent chance of being the mean of a three-item sample.

Table I.4: The Relative Frequency of Means of Samples of Three From a Population of Digits Between 0 and 9

Interval	Density
0.000	0.1
0.333	0.3
0.667	0.6
1.000	1.0
1.333	1.5
1.667	2.1
2.000	2.8
2.333	3.6
2.667	4.5
3.000	5.5
3.333	6.3
3.667	6.9
4.000	7.3
4.333	7.5
4.667	7.5
5.000	7.3
5.333	6.9
5.667	6.3
6.000	5.5
6.333	4.5
6.667	3.6
7.000	2.8
7.333	2.1
7.667	1.5
8.000	1.0
8.333	0.6
8.667	0.3
9.000	0.1

Figure I.4: The Relative Frequency of Means of Samples of Three From a Population of Digits Between 0 and 9

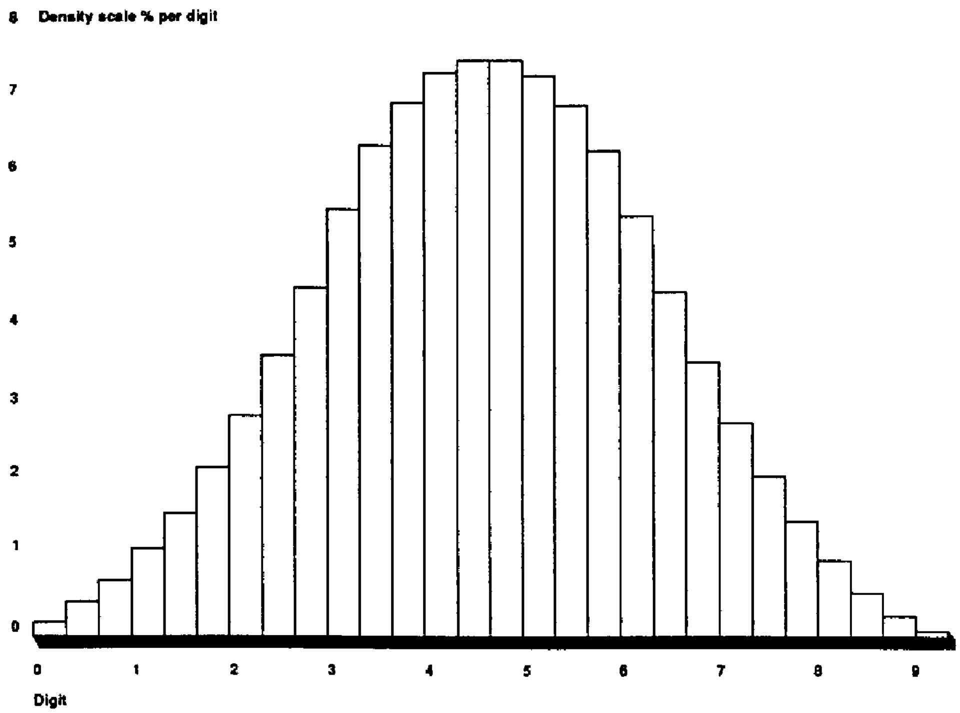
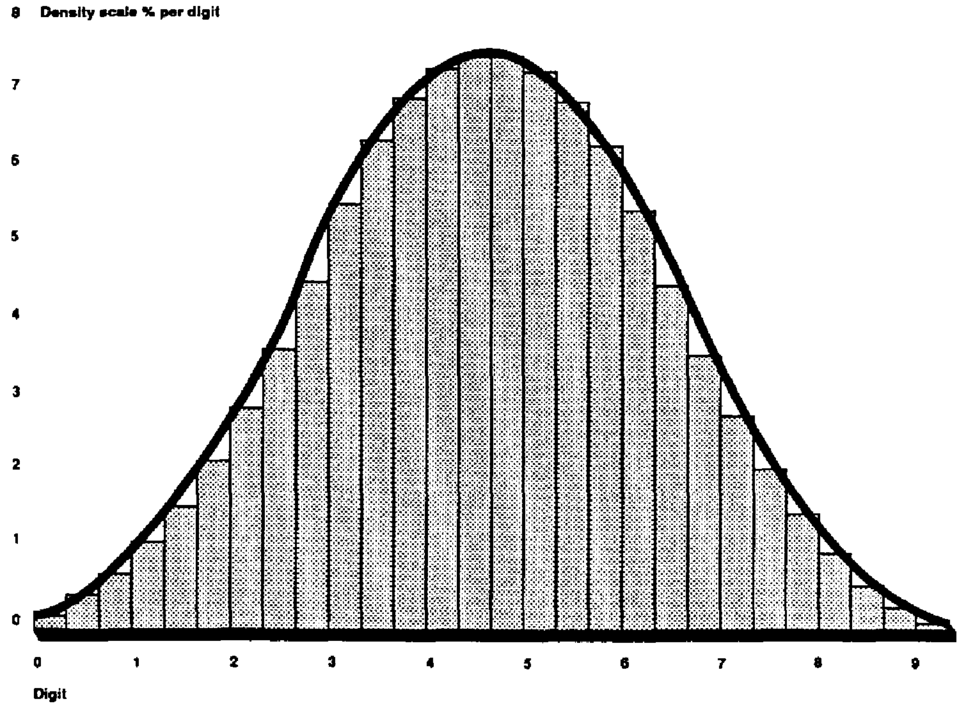


Figure I.5: The Smoothed Curve of the Relative Frequency of Means of Samples of Three From a Population of Digits Between 0 and 9



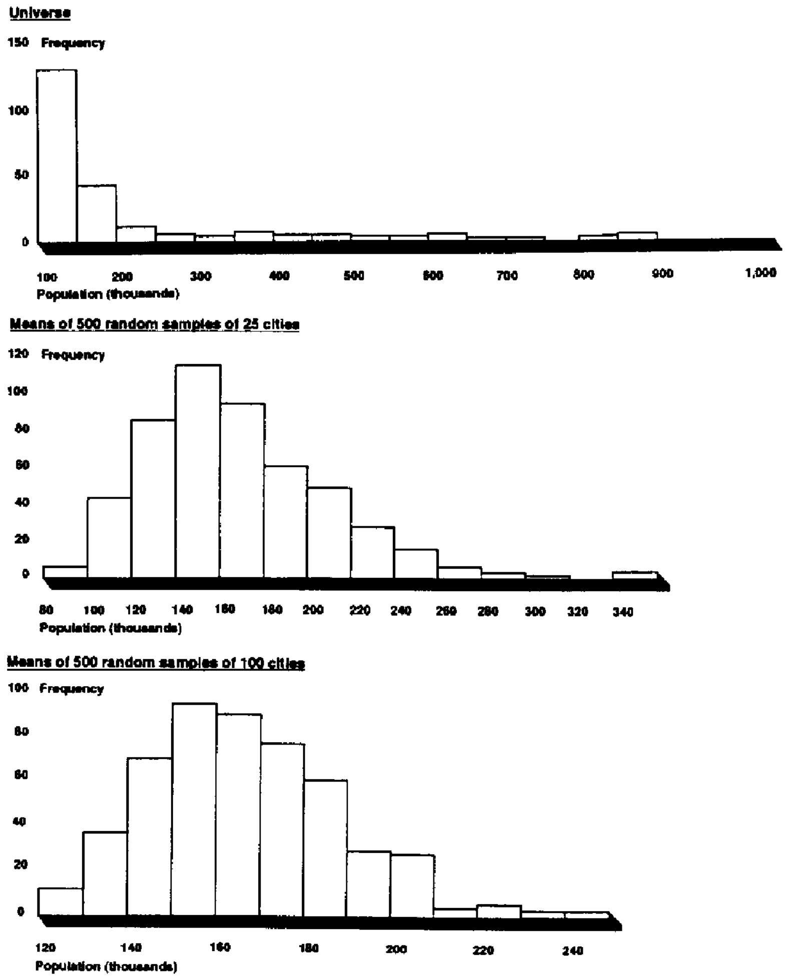
As a sample size increases, the approximating smooth curve, analogous to the one in figure I.5, approaches the normal distribution density curve with a mean of 4.5 and a standard deviation of 2.87 divided by the square root of the sample size. This tendency of the histogram of the means of all simple random samples from any population to approximate a normal curve, if the sample size is sufficiently large, is called the central limit theorem.

To illustrate the power of the central limit theorem, we can note that the normal curve has the following two properties: 68 percent of the area under the curve is within one standard deviation of the mean (or the center of the curve), and 95 percent is within two standard deviations of the mean.

Further, note that the normal curve that approximates the sample mean histogram for a large sample size has a standard deviation of 2.87 divided by the square root of the sample size. Here, 2.87 is the standard deviation of the population histogram in figure I.2. Hence, when the sample size is large, the standard deviation of the sample mean curve (analogous to figure I.5) is quite small; this says that the sample mean of a large random sample will almost certainly be quite close to the mean of the approximating normal curve, which is also the true population mean. For example, if we had a sample size of 36, and the normal curve is therefore a very good approximation of the sample mean curve in the current example, then 95 percent of the possible samples of size 36 will have sample means of between 3.543 and 5.457— that is, $4.5 - (2)(2.87)/\sqrt{36}$ and $4.5 + (2)(2.87)/\sqrt{36}$.

To further illustrate the central limit theorem behavior, we can consider the frequency distribution of 228 cities whose populations were more than 50,000 in 1950, as shown in figure I.6. (The four largest cities have been excluded.) This frequency distribution is definitely not symmetrical. Technically, it is called a reverse J-shaped distribution. The smallest class, cities with populations of 50,000 to 100,000, contains more cities than all the other classes combined. One city with 1,850,000 inhabitants is not even shown because it would require making the horizontal axis twice as long.

Figure I.6: Frequency Distributions of the Means of Samples Drawn From 228 U.S. Cities With Populations Greater Than 50,000 in 1950[a]



[a]The four largest cities are excluded.

Source: Adapted from G. W. Snedecor and W. G. Cochran, Statistical Methods, 7th ed. (Ames, Iowa: Iowa State University Press, 1980).

What happens if we take 500 random samples, each consisting of 25 cities, from this population of 228 cities and prepare a histogram of the distribution of the sample means? (See figure I.6.) We note that the distribution, although by no means symmetrical, is approaching symmetry about the true population mean. This is remarkable, considering the shape of the original distribution.

If we take random samples of 100 cities each and prepare a frequency distribution of the means (as in the third graph in figure I.6), the distribution shows some additional improvement in the direction of symmetry, and sample means cluster more closely about the true population mean, although the distribution is certainly not normal. If the city with 1,850,000 persons had been removed from the population, the distribution of the sample means for a sample of 100 would be more nearly normal. Thus, regardless of the shape of the original population distribution, the distribution of sample means approaches normality as the sample size increases.

Calculating the Standard Deviation and Sampling Error

The standard deviation of a distribution of sample means represents the expected differences between the true population mean and the sample means because of variation from sample to sample. It is called the standard error in order to distinguish it from the standard deviation of the individual population values. (Note that this is strictly a mathematical concept. It does not imply that there has been a human mistake or mechanical failure in some operation.) The standard error is related to the variation in the population distribution. To show how this relationship works, we need to develop further the concept of the standard deviation.

As we stated in chapter 4, the standard deviation is merely a numerical measurement of the dispersion of a group of values about their arithmetic mean.

Although other measures of dispersion exist, the standard deviation has the advantage that it can be manipulated arithmetically—that is, multiplied and divided (but in general not added or subtracted).

Using the population of 100 small purchases from our example in chapter 4, we can calculate a population standard deviation by use of the formula in chapter 2. The 100 small purchase orders are given in table I.5.

Table I.5: The Amounts of 100 Small Purchases

Number	Amount
1	$157
2	147
3	259
4	152
5	144
6	187
7	192
8	189
9	165
10	88
11	166
12	192
13	190
14	185
15	164
16	279
17	230
18	150
19	297
20	199
21	187
22	137
23	261

(continued)

Number	Amount
24	132
25	159
26	218
27	200
28	134
29	259
30	125
31	204
32	177
33	172
34	194
35	197
36	195
37	277
38	74
39	217
40	215
41	237
42	184
43	176
44	169
45	184
46	142
47	177
48	80
49	191
50	231
51	178
52	125
53	172
54	159
55	225
56	241

(continued)

Number	Amount
57	177
58	232
59	170
60	123
61	175
62	169
63	192
64	193
65	233
66	198
67	205
68	226
69	236
70	250
71	203
72	161
73	152
74	202
75	143
76	169
77	218
78	89
79	248
80	160
81	175
82	224
83	159
84	138
85	158
86	194
87	228
88	187
89	177

(continued)

Number	Amount
90	190
91	178
92	164
93	135
94	147
95	96
96	187
97	256
98	163
99	181
100	56

Using the formulas we discussed in chapter 3, we can calculate the population mean of $182.57 and the population standard deviation of $44.66.

The standard deviation of the sample means, or the standard error, can be calculated easily from the population standard deviation. The computation (assuming sample size equals 30) is simply dividing the population standard deviation by the square root of the sample size. Therefore, in this case we have as the value for the standard error $44.66 divided by the square root of 30, or $8.15.

Even before a large sample is drawn, we can say that the chances are about 2 in 3, or the probability is approximately 68 percent, that the sample mean will lie within one standard error of the true mean. Likewise, we can say before the sample is drawn that the chances are approximately 19 in 20, or the probability is approximately 95 percent, that the sample mean will lie within two standard errors of the true mean. In general, these statements will be true regardless of the shape of the population distribution, if the sample size is sufficiently large.

The question that may be asked is, What good is this? The evaluators do not know what the true mean and standard deviation are. If they did, they would not have to sample. The answer is that we take a sample from the population of values and calculate the sample mean and sample standard deviation. We then let the sample standard deviation represent the population standard deviation to calculate the standard error. In chapter 4, we calculated a standard deviation of $48.71 for our sample of 30 small purchase orders. Thus, the estimated standard error is found by dividing the $48.71 by the square root of 30, or $8.89. The standard error of the mean, estimated from the sample, is $8.89. Note that the size of the standard error is inversely proportional to the square root of the sample size. The larger the sample, the smaller the standard error for a given standard deviation.

Once we have computed the standard error, we can say the estimated mean calculated in chapter 4 is $182.30 with a standard error of $8.89. How closely does this estimate the true but unknown mean? We know that if a large number of samples is taken, approximately 68 percent of the sample means will be within one standard error of the true mean. Thus, the probability is approximately 68 percent that our sample mean is within one standard error of the true but unknown mean. From these statements, we can infer that the true but unknown mean is within one standard error of the sample mean approximately 68 percent of the time.

The sample gives us an estimate of the true but unknown mean, and the standard error tells us how precise the estimate is. Although we can say before the sample is drawn that the probability is 95 percent that a sample mean will be within two standard errors of the true mean, we cannot say the probability is 95 percent that the true mean is within two standard errors of the sample mean after drawing the sample and computing the estimates. Either the true but unknown mean is included within the interval given above, in which case the probability is one, or the true mean is not included in the interval, in which case the probability is zero.

Statements such as "68 percent of the sample means are within one standard error on either side of the true mean" and "95 percent of the sample means are within two standard errors on either side of the true mean" are known as confidence statements. The confidence level to be used is set by management or the evaluators, and they base it on the risk they are willing to take that the sample estimate may miss the mark. For example, if management is willing to take a 5-percent risk that the estimate is not correct, the confidence level should be set at 95 percent. If management sets the risk of the estimate not being correct at 1 percent, the confidence level should be set at 99 percent.

The person who does the sampling then selects the proper t factor for the specified confidence level and multiplies it by the standard error to get the maximum possible difference between the sample mean and the true but unknown mean for that confidence level (also called sampling error or precision). (Some t factors are shown in table I.6.) The t factors are based on the normal distribution, which describes the variability in sample means of large size samples. For this example, at 68-percent confidence, the t factor equals 1, so the sampling error is $8.89. At 95-percent confidence, the sampling error equals 1.96 times $8.89, or $17.43.

Table I.6: Table of t Factors

Confidence level as percent	Multiplier for standard error (t factor)
50	0.6745
68	1.000
80	1.282
90	1.645
95	1.960 (or 2)
98	2.326
99	2.578
99.5	2.810
99.73	3.000
99.9	3.291

A Comprehensive Description of Sampling Procedures

This appendix, an extension of the material in chapter 7, provides a comprehensive description of sampling procedures, problems that may confront the sampler, and methods for overcoming these problems. We have attempted to make this material as comprehensive as possible, but it obviously cannot cover every situation that may be encountered nor does it cover every method of random selection.

Random Number Sampling

As mentioned in chapter 7, random number sampling is, in its simplest form, a procedure by which a quantity of random numbers equal to a specified sample size is selected from a population of random digits, called a table of random digits, and then matched against the serial numbers, stock numbers, transaction numbers, or the like that have been assigned to the sampling units in the population of interest. The sampling units having numbers that correspond to the selected random numbers constitute the sample.

Although the description of the procedures to be used in various situations makes the use of random number sampling seem tedious, the work can be done quite quickly by relatively inexperienced personnel once they understand the procedures. The major pitfalls are the failure to define the population of interest correctly and to ensure that the sample is drawn from the entire population.

A table of random digits is a population of thousands of digits from 0 through 9 and is generated by electronic procedures designed to ensure that all the digits have an equal probability of being generated, regardless of which digit was previously generated. The random digits are printed in tables in the same order in which they are generated. Thus, we can use the tables to select any required sample with complete assurance that the sample was drawn at random.

The best known tables of random digits are

- Interstate Commerce Commission (ICC), Bureau of Transport Economics and Statistics, Table of 105,000 Random Decimal Digits (Washington, D.C.: 1949), and
- Rand Corp., A Million Random Digits (New York: Glencoe Free Press, 1955).

Table II.1 contains excerpts from the ICC publication. In the original table, the digits are printed in groups of 5 with 14 columns to the page, and each line has a unique number. In the whole table, the line numbers go from 1 to 1,500. Thus, any group of five digits can be located by the line number and the column number. The digits are printed in groups of five merely to make the table easier to read. The evaluator may read down the columns or across the rows, whichever is more convenient. If the number to be selected consists of fewer than five digits, the evaluator may read from either the left or right edge of the group of five. If the numbers to be selected consist of more than five digits, the digits in one group of five can be combined with the digits from one or more adjacent groups to form a number with the required quantity of digits. It is extremely important that the decisions about the combinations of digits, about whether to read from the left or the right, and about what to do when the bottom of the page is reached be made before the selection process is begun and is followed consistently throughout.

Table II.1: Random Decimal Digits

Line/Col	(1)	(2)	(3)	(4)	(5)	(6)
801	33993	51249	78123	16507	57399	77922
802	39041	05779	74278	75301	01779	60768
803	56011	26839	38501	03321	43259	73148
804	07397	95853	45764	43803	76659	57736
805	74998	53337	13860	89430	95825	65893
806	59572	95893	69765	43597	90570	60909
807	74645	13940	28640	00127	04261	17650
808	42765	23855	38451	11462	32671	52126
809	66561	56130	30356	54034	53996	98874
810	50670	13172	31460	20224	34293	59458
811	53971	08701	38356	36149	10891	05178
812	47177	03085	37432	94053	87057	61859
813	41494	89270	48063	12253	00383	96010
814	07409	32874	03514	84943	74421	86708
815	03097	12212	43093	46224	14431	15065
816	34722	88896	59205	18004	96431	41366
817	48117	83879	52509	29339	87735	97499
818	14628	89161	66972	19180	40852	91738
819	61512	79376	88184	29415	50716	93393
820	99954	55656	01946	57035	64418	29700
821	61455	28229	82511	11622	60786	18442
822	10398	50239	70191	37585	98373	04651
823	59075	81492	40669	16391	12148	38538
824	91497	76797	82557	55301	61570	69577
825	74619	62316	80041	53053	81252	32739

(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
38198	63494	00278	30782	33119	64943	17239	69020
22023	07510	67883	55288	67391	54188	31913	29733
43615	49093	91641	77179	50837	48734	85187	41210
44801	45623	23714	69657	87971	24757	94493	78723
96572	73975	19577	87947	23962	78235	64839	73456
06478	77692	30911	08272	81887	57749	02952	51524
34050	78788	57948	36189	88382	72324	59253	30258
23800	02691	57034	34532	19711	71567	90495	55980
78001	29707	91938	72016	16429	69726	41990	33673
24410	01366	68825	22798	52873	18370	15577	63271
55653	31553	20037	39346	28591	13505	04446	92130
97943	81113	62161	11369	54419	58886	89956	12857
41457	54657	46881	75255	29242	07537	53186	95083
34267	66071	62262	99391	61245	95839	75203	93984
18267	60039	62089	38572	70988	17279	05469	28591
50982	92400	59369	43605	26404	04176	05106	08366
42848	81449	80024	81312	59469	91169	70851	90165
23920	75518	32041	13411	61334	52386	33582	72143
96220	82277	64510	43374	09107	28813	41848	08813
99242	42586	11583	82768	44966	39192	82144	05810
36508	98936	19050	57242	33045	54278	21720	87812
67804	84062	27380	75486	63171	24529	60070	66939
73873	68596	25538	83646	61066	45210	24182	18687
23301	31921	09862	73089	69329	41916	41165	34503
65201	92165	93792	30912	59105	76944	70998	00317

Line/Col	(1)	(2)	(3)	(4)	(5)	(6)
826	12536	80792	44581	12616	49740	86946
827	10246	49556	07610	59950	34387	70013
828	92506	24397	19145	24185	24479	70118
829	65745	27223	22831	39446	65808	95534
830	01707	04494	48168	58480	74983	63091
831	66959	80109	88908	38757	80716	36340
832	79278	02746	50718	90196	28394	82035
833	11343	22312	41379	22797	71703	78729
834	40415	10553	65932	34938	43977	39262
835	72774	25480	30264	08291	93796	22281
836	75886	86543	47020	14493	38363	64238
837	64628	20234	07967	46676	42907	60909
838	45905	77701	98976	70056	80502	68650
839	77691	00408	64191	11006	39212	26862
840	39172	12825	43379	57590	45307	72206
841	67120	01558	99762	79752	17139	52265
842	88264	85390	92841	63811	64423	50910
843	78097	59495	45090	74592	47474	56157
844	41888	69798	82296	09312	04150	07616
845	46610	07254	28714	18244	53214	39560
846	29213	42101	25089	11881	77558	72738
847	38601	25735	04726	36544	67842	93937
848	92207	10011	64210	77096	00011	79218
849	30610	13236	33241	68731	30955	40587
850	74544	72806	62236	65685	37996	00377

(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
41819	85104	25705	92481	95287	61769	29390	05764
64460	96719	43056	24268	23303	19863	43644	76986
42708	54311	95989	08402	77608	98356	47034	01635
03348	11435	24166	62726	99878	59302	81164	08010
81027	72579	67249	48089	34219	71727	86665	94975
30082	43295	37551	18531	43903	94975	31049	19033
03255	39574	41483	12450	32494	65192	54772	97431
65082	57759	79579	41516	46248	37348	34631	88164
95828	98617	27401	50226	17322	44024	23133	57899
51434	66771	20118	00502	07738	31841	90200	46348
16322	45503	90723	35607	43715	85751	15888	80645
73293	38588	31035	12226	37746	45008	43271	32015
24469	15574	40018	90057	96540	47174	03943	37553
99863	58155	66052	96864	61790	11064	49308	94510
53283	75882	93451	44830	06300	45456	49567	51673
97997	66806	55559	62043	51324	32423	88325	99634
38189	88183	56625	22910	58250	70491	71111	37202
88287	47032	66341	38328	70538	91105	12056	36125
34572	83202	58691	27354	37015	11278	49697	65667
68753	16825	48639	38228	95166	53649	05071	26894
57234	28458	74313	29665	97366	94714	48704	07033
68745	62979	97750	28293	75851	08362	71546	17993
52123	29841	76145	82364	55774	15462	44555	26844
45206	11949	28295	12666	98479	82498	49195	46254
59917	91100	07993	15046	51303	19515	25055	56386

Source: Interstate Commerce Commission, Bureau of Transport Economics and Statistics, Table of 105,000 Random Decimal Digits (Washington, D.C.: 1949), p. 17.

Locating a Starting Point in a Random Decimal Digit Table

Take care to avoid starting at the same place each time the table of random digits is used and to avoid a purposive selection of random numbers. There are two basic methods of locating a random starting place in the table.

1. The simplest method is to start at the beginning of the table the first time it is used and mark lightly through each group of digits read. The next time the table is used, start with the first group of digits immediately following the last group lined through. Continue this process until the entire table is used up. At that point, reading can start again at the beginning of the table.

2. More complex is the so-called random stab method, in which the table pages are allowed to fall open and, without looking, the evaluator stabs the pages with a pencil point and begins with the digit closest to it.

A refinement of the random stab method is to locate the starting point by two stabs of the pencil. On the first random stab, read down the column closest to the pencil point, reading either the four lefthand digits or the four righthand digits, until a number between 0001 and 1,500 is reached. This becomes the line number of the starting place. Then allow the table to fall open again at random, make a second random stab, and again read down the column closest to the pencil point, reading either the two lefthand digits or the two righthand digits, until a number between 01 and 14 is reached. This becomes the column number of the starting point.

Determining the Proper Quantity of Digits to Read

Random number sampling can be used most conveniently if the sampling units in the population are already numbered or can be numbered easily. However, actual physical numbering of the sampling units is not necessary if their location can be established by counting (for example, lines in a ledger or folders in a file drawer).

Suppose you want to select a sample of 50 documents from a population of documents numbered from 1 through 360. After locating the starting point in the table and deciding which way to proceed through the table and whether to read the digits from the left or the right, record the first 50 numbers from 001 through 360. Note that because you are selecting three-digit numbers, you must read three digits at a time. The lowest number eligible for inclusion in the sample is the three-digit group 001. Disregard the number 000 and all numbers greater than 360. If you are sampling without replacement—that is, including an item in the sample only once (which is usually the case in GAO work)—do not use random numbers that duplicate a number previously selected. Instead, use the next available random number from 001 through 360.

If the 360 documents in the population are not numbered but a numbering system can be established by counting the documents, almost the same procedure can be used. The only difference is that the random number selected corresponds not to the document's number but to its location in the file, ledger, or list (for example, the 3rd, 8th, or 25th).

To select a sample of 100 items from a population numbered 458 through 15,936, select the first 100 five-digit numbers from 00458 through 15,936. Numbers less than 00458 or from 15,937 to 99,999 and numbers that duplicate previously selected numbers are disregarded.[1]

Special Problems in Random Number Sampling

Certain situations present special problems in random sampling. If these problems are anticipated through careful research into the characteristics of the numbering system, they can usually be resolved with little difficulty.

Gaps in the Numbering System

Occasionally, a numbering system has gaps in it; that is, certain blocks of numbers are not used. When selecting random numbers from a population numbered in such a manner, simply ignore the random numbers that correspond to the gaps. These are equivalent to out-of-range numbers.

For example, assume that a sample of 20 documents is to be drawn from a population of documents numbered from 1 to 95. However, the numbers 1 through 9, 25 through 29, and 40 through 49 are assigned to a class of documents not defined as being in our population. For all practical purposes, this population really consists of documents numbered from 10 through 24, 30 through 39, and 50 through 95.

Using the GAO random number generator, the following numbers are obtained: 43 (ineligible), 57,

[1]In the environment that GAO works in today, we usually select random numbers by using the computer. The random number generator used in programs such as SPSS, SAS, and DYL-AUDIT, as well as the GAO written random number generator, are approved for use in GAO jobs.

67, 82, 48 (ineligible), 20, 42 (ineligible), 8 (ineligible), 4 (ineligible), 73, 1 (ineligible), 19, 60, 27 (ineligible), 55, 28 (ineligible), 34, 22, 93, 69, 63, 83, 89, 87, 72, 53, 45 (ineligible), and 23. Therefore, the sample would consist of documents 19, 20, 22, 23, 34, 53, 55, 57, 60, 63, 67, 69, 72, 73, 82, 83, 87, 89, 93, and 94.

Caution should be taken in that gaps in the numbering system sometimes indicate that various groups of items have been assigned different blocks of numbers because the characteristics of the groups have important differences on the variable of interest. If so, it might be advisable to define each group as a separate population. Always investigate this possibility before proceeding as if there were only a single population.

Ineligible Items

Occasionally, certain items not eligible for inclusion in the sample are not identified until they have actually been examined. Some of the random numbers may correspond to documents that have been voided, to inventory items that are no longer stocked or otherwise ineligible, and the like. Or certain types of items or entries may not be of interest. For example, payment and receipt entries may be intermingled when only payments are to be sampled.

If you were unaware that certain items would be ineligible for inclusion in the sample, you would use up the random numbers before finding a sufficient quantity of eligible sampling units. To obtain the specified sample size, select additional random numbers, plus some extras to allow for additional ineligible items.

However, the most efficient approach to this problem is to estimate in advance the proportion of usable items, if not already known, by scanning a list of the population items (if available), taking a small preliminary sample of items, or questioning agency officials. The estimated proportion of eligible items is then divided into the required sample size to determine the quantity of random numbers to be selected; that is, the required quantity of random numbers equals the specified sample size divided by the proportion of usable items.

Assume that in a review of travel vouchers, a sample of 300 vouchers involving reimbursement for the use of personally owned vehicles will be required. From a small preliminary sample, the evaluators estimate that the proportion of vouchers with mileage claims is about 75 percent, or 0.75. The required quantity of random numbers is calculated as 300 divided by 0.75, or 400.

After the items corresponding to the random numbers have been examined, the actual sample of eligible items may differ from the specified sample size. A rule of thumb is that differences of less than 10 percent of the specified sample size can be ignored. If the difference is 10 percent or greater, compute the quantity of additional random numbers to be selected as follows:

1. Determine the actual proportion of eligible items in the first sample by dividing the quantity of random numbers selected into the quantity of eligible items found.

2. Divide this proportion into the quantity of additional eligible items needed to determine the quantity of additional random numbers required.

If, in the travel vouchers example, the first sampling operation found only 220 vouchers containing mileage claims among the 400 vouchers examined, the proportion of eligible documents would be 220 divided by 400, or 0.55. Since 80 additional vouchers with mileage claims would be needed to obtain a sample of 300 eligible documents, the quantity of additional random numbers to be selected would be 80 divided by 0.55, or 146.

If the sample of eligible items is larger than required, the sample size can be reduced by using one of the techniques described in the last section of this appendix.

Selecting Random Letters or Months

Sometimes it is necessary to select a series of random letters or random months to draw a random sample. Some publications (such as Arkin, 1984) contain tables of random letters and random months. However, if such tables are unavailable, the problem of selecting a group of random letters can be easily resolved by selecting a group of random numbers from 1 to 26 and assigning the letters of the alphabet that correspond to the numbers selected, such as A for 1, B for 2, and C for 3. Similarly, random months can be selected by selecting random numbers from 1 to 12 and assigning the months corresponding to the numbers selected.

Compound Numbering Systems

Sometimes numbering systems use a letter as a prefix or suffix to the digits in a number. This is referred to as a compound numbering system, of which there are two types: the quantity of items is the same for each letter or the quantity of items differs for the various letters.

Same quantity of items for each letter used. Following the procedures below will produce an unduplicated random sample drawn from a single population. Assume that a sample of 20 items is required from a population numbered as follows:

- A-0001 to A-5,000
- B-0001 to B-5,000
- C-0001 to C-5,000 through
- Z-0001 to Z-5,000.

Thus, there are 5,000 items for each letter from A to Z.

The first step is to select 20 random numbers from 0001 to 5,000, plus some extras. Record the random numbers, including the extras, in the order in which they were selected, keeping the extras separate. Duplicate numbers should not be eliminated at this point. The second step is to select 20 random letters plus the same number of extra letters as numbers in step one from A to Z. Do not eliminate duplicates. Assume that a selection of 25 four-digit random numbers and 25 random letters yields the results shown in table II.2.

Table II.2: Selecting Random Numbers and Random Letters in Compound Numbering Systems With the Same Quantity of Items for Each Letter[a]

Order of selection	Random letter	Order of selection	Random number
1	M	1	3,284
2	U	2	1,224
3	J	3	0199
4	X	4	0578
5	Y	5	1,240
6	K	6	0750
7	S	7	0994
8	O	8	2,055
9	Y	9	4,038
10	A	10	4,976
11	Y	11	4,815
12	K	12	2,751
13	K	13	1,946
14	F	14	2,814
15	N	15	2,055
16	P	16	1,944
17	Y	17	1,240
18	R	18	4,684
19	F	19	1,353
20	O	20	2,021
	(extras)		(extras)
21	N	21	1,959
22	O	22	1,644
23	N	23	4,768
24	L	24	3,612
25	V	25	1,347

[a]Although F, K, N, O, and Y and 1,240 and 2,055 appear more than once, they are not eliminated at this step.

The random numbers and letters shown in table II.2 were deliberately chosen to illustrate the process, and the quantity of duplicates has been exaggerated.

Normally, it would not be necessary to select a quantity of extras equal to 25 percent of the original sample. The quantity of extras needed depends on the anticipated number of ineligible items (if any) that may be found in the sample and the anticipated number of nonresponses in a personal interview or questionnaire survey for which evaluators may want to substitute other randomly selected sampling units. Normally, a quantity of extras equal to 10 percent of the original sample should be enough.

The third step is to match numbers and letters, keeping both in the original order of selection. The results of the matching process are illustrated in table II.3.

Table II.3: Matching Random Numbers and Random Letters in Compound Numbering System With the Same Quantity of Items for Each Letter

Order of selection	Letter and number
1	M-3,284
2	U-1,224
3	J-0199
4	X-0578
5	Y-1,240
6	K-0750
7	S-0994
8	O-2,055
9	Y-4,038
10	A-4,976
11	Y-4,815
12	K-2,751
13	K-1,946
14	F-2,814
15	N-2,055
16	P-1,944
17	Y-1,240
18	R-4,684
19	F-1,353
20	O-2,021
	(extras)
21	N-1,959
22	O-1,644
23	N-4,768
24	L-3,612
25	V-1,347

In the fourth step, look at the 20 number-letter combinations and eliminate any duplicates. Use the extra number-letter combinations, in the original order of selection, to replace the duplicates. Insert the replacements into the original group. The sorted list of random number-letter combinations is shown in table II.4. Note that extra N-1,959 was used to replace the duplicate Y-1,240. This was the first extra selected. If another duplicate had been discovered, extra O-1,644 would have been used to replace it, and so on.

Table II.4: Sorting Random Number-Letter Combinations in Compound Numbering Systems With the Same Quantity of Items for Each Letter

Order of selection	Letter and number	Remarks
10	A-4,976	
19	F-1,353	
14	F-2,814	
3	J-0199	
6	K-0750	
13	K-1,946	
12	K-2,751	
1	M-3,284	
21	N-1,959	Extra; replaces duplicate Y-1,240
15	N-2,055	
20	O-2,021	
8	O-2,055	
16	P-1,944	
18	R-4,684	
7	S-0994	
2	U-1,224	
4	X-0578	
5	Y-1,240	
17	Y-1,240	Duplicate; replaced by N-1,959
9	Y-4,038	
11	Y-4,815	

Different quantity of items for each letter used. In some instances, the number of items is not the same for two or more letters of the alphabet. This is a more complicated variation of the situation described above. It is important that the procedures described below be followed exactly.

Assume that a sample of 20 items is required from a population numbered as follows:

- A-0001 to A-5,056
- B-0001 to B-5,397
- C-0001 to C-7,409
- D-0001 to D-4,455
- E-0001 to E-4,619
- F-0001 to F-7,691
- G-0001 to G-6,100
- H-0001 to H-5,406

Although the letters here go up only to H, they could go through the entire alphabet or they could start with F and end with Q, for example. The lowest number is 0001, and the highest is 7,691. In some numbering systems, the lowest number for one or more of the groups may be greater than 1; however, the sampling procedure remains the same.

Select the sample as follows.

1. Select 20 random numbers, plus some extras, say 10, between 0001 and 7,691. Record the random numbers, including the extras, in the order of selection, keeping the extras separate. Do not eliminate duplicates.

2. Select 20 random letters, plus 10 extras, between A and H. Record the random letters in the order of selection, keeping the extras separate. Do not eliminate duplicates.

Assume that the selection of random numbers and letters yields the results shown in table II.5. Again, the random numbers and letters shown here were deliberately chosen to illustrate the process. The quantity of duplicates and out-of-bound numbers has been greatly exaggerated. Normally, it is not necessary to select a quantity of extras equal to 50 percent of the original sample; 10 to 15 percent should suffice.

Table II.5: Selecting Random Numbers and Random Letters in Compound Numbering Systems With a Different Quantity of Items for Each Letter[a]

Order of selection	Random letter	Order of selection	Random number
1	G	1	6,385
2	F	2	0718
3	C	3	2,472
4	H	4	1,117
5	E	5	4,236
6	H	6	2,331
7	E	7	3,454
8	C	8	7,101
9	F	9	0742
10	C	10	2,472
11	F	11	0718
12	B	12	5,406
13	C	13	0563
14	G	14	1,438
15	D	15	3,952
16	A	16	7,146
17	A	17	3,158
18	A	18	3,615
19	G	19	2,547
20	B	20	1,992
	(extras)		(extras)
21	E	21	5,493
22	B	22	0317
23	G	23	4,122

(continued)

Order of selection	Random letter	Order of selection	Random number
24	F	24	6,320
25	D	25	6,206
26	B	26	4,071
27	C	27	7,545
28	F	28	3,253
29	B	29	6,817
30	C	30	0598

[a]Although certain numbers and letters appear more than once, they are not eliminated at this step.

The next step is to match the letters and numbers, keeping both in the original order of selection, producing the results shown in table II.6.

Table II.6: Matching Random Numbers and Random Letters In Compound Numbering Systems With a Different Quantity of Items for Each Letter

Order of selection	Letter and number
1	G-6,385
2	F-0718
3	C-2,472
4	H-1,117
5	E-4,236
6	H-2,331
7	F-3,454
8	C-7,101
9	F-0742
10	C-2,472
11	F-0718
12	B-5,406
13	C-0563
14	G-1,438
15	D-3,952
16	A-7,146
17	A-3,158
18	A-3,615
19	G-2,547
20	B-1,992
	(extras)
21	F-5,493
22	B-0317
23	G-4,122
24	F-6,320
25	D-6,206
26	B-4,071
27	C-7,545
28	F-3,253
29	B-6,817

Next, sort the original 20 letter-number combinations into alphabetical-numerical order. (Keep the extras in the original order of selection in a separate group.) Then eliminate two types of combinations: (1) those having no corresponding item number (that is, the out-of-bounds combinations) and (2) those that duplicate a combination previously selected. The eliminated original combinations are replaced by eligible extras in their order of selection.

For example, combination A-7,146 is greater than the highest number for the A group of items. Therefore, replace this out-of-bounds combination by extra B-0317, the first eligible extra in order of selection. Although combination E-5,493 was the first extra selected, it is also out of bounds and cannot be used. Of the original 20 combinations, B-5,406 is also out of bounds and is replaced by extra F-6,320. C-2,472 was selected twice, the third and tenth combinations drawn. The duplicate C-2,472 is eliminated and replaced by extra G-4,122, the third eligible extra. When used as replacements, the extras are put in their proper alphabetical-numerical sequence in the original sample. Continue the process until the required quantity of unduplicated, within-bounds letter-number combinations has been obtained. In this example, all eligible extras except C-0598 are used. (Extras E-5,493, D-6,207, C-7,545, and B-6,817 are out of bounds and cannot be used.) The final list of letter-number combinations appears in table II.7.

Table II.7: Sorting Random Letter-Number Combinations in Compound Numbering Systems With a Different Quantity of Items for Each Letter

Order of selection	Letter and number	Remarks
17	A-3,158	
18	A-3,615	
16	A-7,146	Out-of-bounds; replaced by B-0317
22	B-0317	Extra; replaces A-7,146
20	B-1,992	
26	B-4,071	Extra; replaces F-0718
12	B-5,406	Out-of-bounds; replaced by F-6,320
13	C-0563	
3	C-2,472	
10	C-2,472	Duplicate; replaced by G-4,122
8	C-7,101	
15	D-3,952	
7	E-3,454	
5	E-4,236	
2	F-0718	
11	F-0718	Duplicate; replaced by B-4,071
9	F-0742	
28	F-3,253	Extra; replaces G-6,385
24	F-6,320	Extra; replaces B-5,406
14	G-1,438	
19	G-2,547	
23	G-4,122	Extra; replaces duplicate C-2,472
1	G-6,385	Out-of-bounds; replaced by F-3,253
4	H-1,117	
6	H-2,331	

When items are listed in a book or on a computer printout, a random number sample can be drawn by using both the page number and the line number. The technique is almost the same as that described above for compound numbering systems, except that the prefix is the page number instead of a letter, and the line number is the remaining portion of the number.

For example, assume that the evaluators want to draw a sample of inventory items from an 80-page catalog listing all items in stock. The maximum number of items that can be listed on a page is 156. Each item in the catalog can be identified by a two-part numbering system: the page number (from 1 to 80) and the line number (from 1 to 156). To draw the sample, merely select the required quantity of random numbers between 1 and 156, plus some extras, and the required quantity of random numbers between 1 and 80, plus extras, without eliminating duplicates. Match the two series of numbers in order of selection and sort them into numerical order, then eliminate duplicates, out-of-bounds numbers, and numbers corresponding to ineligible items (if any). The eliminated numbers are replaced by extras in order of selection. The result is a simple random sample from the catalog.

If the items are printed in two or more parallel columns, count the items instead of the line numbers. This procedure can also be extended to more complicated numbering systems. For example, the numbering system may consist of three groups of digits arranged as 10-450-39. The first two digits represent the folio or book number, the next three represent the page number, and the last two represent the line number. Both the number of pages per book and the number of lines per page can vary. For this type of population, first select a sufficient quantity (plus extras) of two-digit numbers for the lines, three-digit numbers for the pages, and two-digit numbers for the book numbers. Match the numbers in order of selection and replace duplicates and out-of-bounds combinations by the extras in order of selection.

Caution should be taken when sampling from books or computer printouts: it is usually incorrect to assume that the same number of eligible cases will be listed on each page.

Periodic Serial Numbering System

Some numbering systems assign numbers serially (1, 2, 3, and so on) to documents for a certain period such as a week or a month. At the beginning of the next time period, the numbering process starts again with 1. This type of system appears to assign the same identification numbers to different items. If, however, the time periods are assigned numbers 1 through 12 for months or 1 through 52 for weeks, this type of numbering system becomes almost identical to the page-number and line-number system described in the previous section. The number of the time period corresponds to the page number, and the serial number corresponds to the line number. Thus, the sample can be drawn by the same procedure used for sampling from a book or computer printout. The number of items should be expected to differ from one time period to another.

Systematic Selection With a Random Start

We explained the theory of systematic selection with a random start in chapter 7. In this section, we present some detailed examples of the procedure, including some unusual situations. To use systematic sampling with a random start, assume that you want to sample 200 items from an unnumbered list of 4,500 items. Divide the number of items, 4,500, by the specified sample size, 200, to get 22.5; then drop the digits to the right of the decimal point, regardless of whether their value is more or less than 0.5. This ensures that the sample is at least the required size. A sampling interval of 22 results in a sample size of 204 or 205, depending on the starting place. An interval of 23 provides a sample of only 194 or 195 items.

Then, using a method of selecting a random number, select a two-digit number between 01 and 22 as the starting point. Assume that this number is 13. Count the items on the list until item 13, which will be the first sample item, is reached. Count off 22 more items and take the 35th item for the 2nd sample item; then count 22 more items and take the 57th item; repeat through the 4,479th item. With this method, the entire sample of 204 items can be selected without bothering to number the items.

Consider a second example. Assume that a sample of 180 items is required from a list of items numbered serially from 41,001 through 44,000. First determine that there are no gaps in the numbering system—that is, that all the numbers between 41,001 and 44,000 have been used. Then subtract the first number used from the last number used and add 1 to the difference to obtain the population size:

Last number	44,000
First number	41,001
Difference	2,999
	+ 1
Population size	3,000

Divide the population size, 3,000, by the specified sample size, 180, to obtain the sampling interval. In this case, the sampling interval will be 16, after dropping the digits to the right of the decimal point.

Select a two-digit random number between 01 and 16. Assume that the random number is 05. This number equals the last two digits of the serial number of the first sample item, so the first sample item is 41,005.

Then add the sampling interval, 16, to the serial number of the first sample item, 41,005, to obtain the serial number of the second sample item, 41,021. Continue adding the sampling interval to the serial number of the item previously selected to obtain the serial number of the sample item, until you obtain a number larger than the last serial number in the population. Thus, the items with serial number 41,005, 41,021, 41,037, and so on through 43,997 will be selected. (The serial numbers of the sample items can be easily determined by using an adding machine with a paper tape and taking subtotals after addition.)

Consider a third example. Assume that records in a single population are maintained in four groups numbered as in table II.8. These records are numbered in a broken series. To select a single systematic sample from the four groups of records, subtract the beginning serial number of each group from the ending serial number and add 1 to the difference to obtain the total number of records in each group. Then add the total numbers of records to obtain the population size for all four groups combined. Assign mentally designated serial numbers to indicate the numerical sequence of each item in the population. An example of the computation is shown in table II.9.

Table II.8: A Population of Records in Four Groups

Group	Assigned serial numbers
A	14,542 through 17,921
B	19,055 through 19,988
C	22,001 through 23,021
D	25,500 through 26,401

Table II.9: Selecting a Single Systematic Sample From Four Groups of Records

Group	Serial number		Number of records	Mentally designated serial number	
	Beginning	Ending		Beginning	Ending
A	14,542	17,921	3,380	1	3,380
B	19,055	19,988	934	3,381	4,314
C	22,001	23,021	1,021	4,315	5,335
D	25,500	26,401	902	5,336	6,237

Next, divide the population size by the required sample size, dropping decimals, to obtain the sampling interval. Suppose that from the population of 6,237 records, a sample of 210 records is wanted. The sampling interval will be 6,237 divided by 210, or 29.7, or 29.

Select a two-digit random number between 01 and 29. Assume the number is 03. This means that the third record will be the first sample item. The serial number of the first sample item is determined by subtracting 1 from the first number in group A and adding the random number 03 to the difference. The number of the first record selected for the sample will be 14,544 (14,542 - 1 + 3 = 14,544).

Determine the serial numbers of the second and successive sample items by adding the sampling interval, 29, to the serial number of each record previously selected. Thus, records with serial numbers 14,544, 14,573, 14,602, 14,631, and so on are selected from group A. Then determine the serial number of the last record to be selected from group A and the first record from group B by following the steps described below.

1. Subtract the serial number of the first record selected from the group from the highest serial

number assigned to that group to obtain the balance of the group. For example:

Highest serial number assigned to group A	17,921
Less serial number of first record selected from group A	-14,544
Balance of the group	3,377

2. Divide the balance of the group by the sampling interval and subtract the remainder from the balance of the group to obtain the group difference—that is, the difference between the serial numbers of the first and last records selected from the group. For example, balance of group A (3,377) divided by sampling interval (29) equals 116 with a remainder of 13

Balance of the group	3,377
Less remainder	- 13
Group difference	3,364

3. Add the group difference to the serial number of the first record selected from the group to obtain the serial number of the last record to be selected from the group.

Serial number of the first sample record from group A	14,544
Plus group difference	+ 3,364
Serial number of last sample record from group A	17,908

4. Subtract the remainder, obtained in step 2 above, from the sampling interval to determine the sequence

of the first record to be selected from the next group. For example:

Sampling interval	29
Less remainder from group A	– 13
Sequence of first sampling record to be selected from group B	16

5. Subtract 1 from the lowest serial number assigned to the next group and add the sequence obtained in step 4 to determine the serial number of the first sample record to be selected from the next group. For example:

Lowest serial number in group B	19,055
Less 1	– 1
Difference	19,054
Plus sequence of first sample record	+ 16
Serial number of first sample record to be selected from group B	19,070

Use this procedure when going from one group to another, continuing it through all groups until the entire sample has been selected. The actual sample size will be 215 because the calculated interval of 29.7 was rounded downward to 29.

Sampling by Measurement

If the sampling units are equal-width documents such as index cards, punch cards, or sheets of paper, a quick systematic sample can be obtained by measurement. This method should be used with caution if there is any likelihood that the findings will be controversial, because it is impossible to document the sampling unit that was closest to the ruler's mark.

Simply measure the total length of the file of records and divide this by the desired sample size to obtain a sampling interval in inches or fractions of inches. For example, assume that you want to select a sample of 160 cards from two file drawers, each measuring 2-1/2 feet in length. The steps are as follows.

1. Convert the total length of the two file drawers, or 2-1/2 ft. + 2-1/2 ft. = 5 ft., into inches by multiplying by 12. In this case 5 x 12 = 60 in.

2.To obtain the required sampling interval, divide the total length of the files by the required sample size: 60/160 = 3/8 in.

3. Select the sample by laying a ruler on top of the file and selecting the cards that are opposite each 3/8-inch mark on the ruler. To obtain a random start, place the end of the ruler at a randomly selected card between the beginning of the file and a point that is 3/8 inch along the file.

Sometimes the sampling interval obtained by dividing the sample size into the total length of the population will not coincide with one of the fractional parts of an inch marked on a ruler. If this happens, round the sampling interval downward to coincide with the closest marking. For example, if the quotient obtained by dividing the required sample size into the length of the population were 0.65 inch, the sampling interval should be 5/8 inch (0.625 inch). If the quotient were 1.4 inches, the sampling interval should be 1-3/8 inches (1.375 inches). This sample selection method is much easier to use if a measuring instrument in centimeters and millimeters is available.

When the sample is to be drawn from a large number of items, selection by measuring, for all practical purposes, is equivalent to systematic selection by counting. However, selection by measurement cannot be used if the sampling units are of varying widths, because the thicker items will have a greater probability of being selected.

A similar problem arises in systematic selection by counting if the same sampling unit has several cards, folders, and the like or appears more than once on a list. When this occurs, all the cards, lists, and so on for the sampling unit must be considered a single sampling unit and counted as such; otherwise, these sampling units will have a greater probability of being selected than those listed once or having only one folder.

If it is not possible or efficient to combine several lists before sampling, alternative procedures may be used. One is to sample from each of the lists or sources available. For all selected units, check each list to see on how many of the lists they are found. Then subsample inversely to the number of times found; in other words, take a sample of half of the items on two lists, one third of the items on three lists, and so on. If it is impossible to determine how many lists a unit is on until the data are gathered, the data will have to be weighted to develop unbiased estimates.[2]

Expanding a Systematic Sample

Sometimes it may be necessary to increase the size of a systematic sample. The simplest method is to select a quantity of additional random starting points between 1 and the original sampling interval such that when the items selected by taking every nth item are selected, the result will be a total sample size approximately equal to the specified sample size.

[2]We are grateful to Seymour Sudman of the University of Illinois, Champaign, Ill., for pointing this out.

Another method is, in effect, to redefine the population by excluding the items selected in the original systematic sample. Using the procedures described above, select from the redefined population a supplementary systematic sample that is approximately equal in size to the number of additional sampling units required. To make sure that none of the sampling units selected for the original sample are counted when locating the random starting point and the additional sampling units, identify them by check marks, paper clips, or the like. After the supplementary sample has been selected, combine it with the original sample and consider it to be a single sample in which each unit's probability of being selected is equal to the final sample size divided by the population (n/N).

Assume that the evaluators have selected a preliminary sample of 30 items from a population of 9,000 items by starting with the 2nd sampling unit (selected randomly) and taking every 300th item thereafter. Identify the sample items by check marks in the list from which they were selected. If after analyzing the results of the preliminary sample you decide that the final sample should have 400 items, follow these three steps to obtain the 370 additional items.

1. Redefine the population to exclude items selected for the preliminary sample and subtract the preliminary sample size, 30, from the population size, 9,000, to obtain the size of the redefined population, 8,970 items.

2. Divide the number of items required for the supplementary sample, 370, into the redefined population size, 8,970, to obtain the sampling interval for the supplementary sample, 24.

3. Select a random number between 1 and 24 to determine the starting point for the systematic sample

from the redefined population. Whatever the starting point may be, do not count, but merely skip over, original sample items (2, 302, 602, and so on) in selecting the supplementary sample.

If it is not practical to identify on the list the items selected for the preliminary sample, a third method can be used to expand a systematic sample. Subtract the number of items in the original sample from the population size and divide the difference by the number of items required for the supplementary sample to obtain the sampling interval. Use this interval to select a systematic sample from the original population, proceeding as if the original sample had not been selected. The second sample is combined with the original sample, and items from the second sample that duplicate items in the preliminary sample are eliminated.

Assume the same situation that existed in the previous example, except that when actually going through the population, you are unable to identify readily the items selected in the preliminary sample. Follow these four steps.

1. Calculate the sampling interval for the supplementary sample as 1 in 24, exactly as in the previous example.

2. Select a systematic sample, using the sampling interval of 1 in 24, from the entire original population of 9,000 items, as if the preliminary sample had not been selected. The second sample will contain about 375 items.

3. Compare the items in the supplementary and preliminary samples and eliminate from the supplementary sample the items that are duplicates of items selected in the preliminary sample. (About 1.3 percent of the items in the supplementary sample will be eliminated.)

4. Combine the two samples to obtain an unduplicated sample of about 400 items, which is the sample size required.

Reducing the Size of a Sample

Sometimes it is necessary to reduce the size of, or "thin out," a sample already selected. Some possible reasons for this follow.

1. If the sample was drawn by systematic selection with a random start, the population size may be larger than originally estimated and, as a result, the sample may be larger than required.

2. You may discover that it takes longer than anticipated to examine a sample item and examining all the items selected would take more time than justified by the objective of the job.

3. You may have only a single opportunity to draw the sample. To ensure that the final sample will have sufficient items, you may select many more items than it was estimated the final sample would require. Draw a smaller sample, or subsample, from this large sample for the preliminary sample. You may also be able to use an additional subsample of the large sample as a supplementary sample to achieve the final required sample size.

Basically, thinning a sample is drawing a randomly selected subsample of items from a sample that has already been selected. All items in the original sample must have equal opportunity of being selected in the subsample. If they do not, the subsample will not be representative of the population.

One of the simplest methods of thinning a sample is to use systematic selection with a random start to select either the items to be eliminated or those to be retained. This method will work regardless of the procedure used to select the original sample and, in general, regardless of the sequence of the sample items, unless the evaluators suspect that the sequence may cause the characteristic being measured to recur at regular intervals. This can sometimes be detected by inspecting a list of the sample items and the corresponding values. If this situation is suspected, use random number sampling to thin the sample.

If the sample items have consecutively assigned identification numbers, the sample can be thinned by using sampling that is based on randomly selected combinations of terminal digits (discussed in chapter 7).

If random number sampling was used to draw the original sample and the random numbers are still in the order of selection or can be rearranged into that order, random numbers can be eliminated by beginning with the last one selected and working back until enough random numbers have been eliminated to reduce the sample to the required size. Or start with the first random number selected and, working forward, count out a quantity of random numbers equal to the required sample size. If the random numbers are arranged in any sequence other than the order of selection, this procedure cannot be used, unless they are randomized by some procedure that can be documented. The numbers must be in some random order, although not necessarily in the order of selection.

Random number sampling can also be used to thin a sample. This procedure will work regardless of how the original sample was selected or how the sample items are sequenced. Simply use a random method to select either the items to be retained in the sample or those to be eliminated. The selection of the random numbers will easier if the sample items have been renumbered consecutively from 1 through the last item in the sample.

Details on Stratified and Cluster Sampling

This appendix discusses the computation of estimates, sampling errors, and sample sizes and the allocation of a sample among strata when stratified sampling is used. It briefly describes how to construct stratum boundaries and how to determine the optimum number of strata. It concludes with a brief description of a two-stage cluster sampling problem.

Types of Stratified Sampling

As we noted in chapter 3, in stratified sampling the sample size in each stratum may be proportional to the total number of sampling units in the stratum ("proportional allocation"), or it may be disproportional. Examples of both kinds of allocation are given in this appendix.

In proportional allocation, the proportional relationship between the stratum sample size and the total sample size is the same as that between the stratum population size and the total population size. That is, the sampling fraction (sample size divided by population size) is the same in all strata.

The advantages of proportional allocation over other allocation methods are (1) the formula for allocating the sample to the strata is simple, (2) the formulas for computing estimates are simple, and (3) proportional allocation is intuitively more familiar to those who use the final results, which may prevent them from making gross errors if they attempt to manipulate the sample results arithmetically.

In disproportional allocation, there are three methods of allocating the sample to the strata. The judgmental method is simply based on the evaluator's desire to meet a specific objective, such as doing a 100-percent audit of all high-value transactions and auditing a sample of the remaining transactions or doing a 100-percent audit of the more error-prone transactions, if they can be identified, and auditing a sample of the others.

The other methods of disproportional allocation are known as Neyman allocation and optimum allocation. The Neyman method allocates the sample to each stratum in proportion to the product of the stratum population size and its standard deviation, divided by the sum over all strata of the products of the stratum population sizes and standard deviations. The standard deviations can be estimated from a preliminary sample or from a prior audit or study. The advantage of the Neyman method is that the precision is minimized for a given sample size.

Optimum allocation allocates the sample to strata by taking into account the differences in (1) population sizes, (2) standard deviations, and (3) the costs of collecting data among the various strata. Optimum allocation minimizes the precision for a specific total cost of data collection or, conversely, minimizes the total cost of data collection for a specified precision. A discussion of optimum allocation is beyond the scope of this paper.

Stratified Sampling for Variables

To illustrate proportional, judgmental, and Neyman sample allocation methods and the procedures for computing estimates, precision, and sample sizes, we can consider the following example of stratified sampling for variables.

While reviewing shipping costs at a supply depot, the evaluators suspect that air freight forwarding, the shipping method used, is less economical than direct air carrier. By calculating the costs of several recent shipments from direct air carrier rate schedules, they find that in each case the direct air shipping cost is less than the amount paid to the air freight forwarder. The evaluators decide to estimate, using statistical sampling, the total saving that would have resulted had direct air carrier been used instead of air freight forwarding. The confidence level is set at 95 percent.

The population is defined as all air freight forwarder shipments during the past 3 months. From the depot's file of shipping documents, the information required to compute the cost of shipments by direct air carrier rate schedules is copied into the evaluators' files. This procedure results in 250 records, each representing a single shipment. Using their judgment, the evaluators classify the records into three groups, based on the air freight forwarder shipping costs, which they know before they do any audit work and are shown in table III.1.

Table III.1: Air Freight Forwarder Shipping Costs

Stratum	Number of shipments
Less than $100	150
$100-$499	75
$500 or more	25
Total	**250**

The resulting savings on each shipment are shown in tables III.2 and III.3. (Note that in real life, the savings would not be known until the shipping costs by direct air carrier had been computed. Calculating shipping costs is a very complicated, time-consuming procedure; this is why sampling is necessary.)

Table III.2: Savings From Using Direct Air Shipment Instead of Air Freight Forwarder Classified by Air Freight Forwarder Shipping Costs of Less than $100[a]

Shipment	Savings
1	$21
2	27
3	33
4	44
5	11
6	52
7	23
8	32

(continued)

Shipment	Savings
9	43
10	39
11	23
12	26
13	19
14	24
15	39
16	22
17	35
18	35
19	39
20	34
21	13
22	19
23	4
24	30
25	31
26	16
27	22
28	13
29	46
30	37
31	47
32	37
33	15
34	27
35	10
36	20
37	35
38	33
39	38
40	26
41	13

(continued)

Shipment	Savings
42	31
43	47
44	51
45	29
46	37
47	29
48	25
49	17
50	18
51	29
52	18
53	25
54	28
55	22
56	35
57	33
58	24
59	15
60	31
61	6
62	19
63	16
64	29
65	41
66	24
67	26
68	18
69	30
70	27
71	29
72	42
73	40
74	18

(continued)

Shipment	Savings
75	31
76	17
77	22
78	40
79	38
80	36
81	21
82	22
83	34
84	42
85	39
86	45
87	34
88	15
89	17
90	21
91	38
92	28
93	24
94	42
95	34
96	28
97	26
98	30
99	23
100	34
101	37
102	37
103	38
104	30
105	42
106	34
107	41

(continued)

Shipment	Savings
108	37
109	45
110	44
111	42
112	23
113	30
114	36
115	15
116	39
117	37
118	28
119	30
120	46
121	36
122	33
123	22
124	20
125	33
126	23
127	45
128	26
129	50
130	0
131	28
132	25
133	17
134	10
135	33
136	44
137	6
138	24
139	22
140	36

(continued)

Shipment	Savings
141	24
142	55
143	25
144	26
145	33
146	43
147	23
148	15
149	30
150	35

[a]In real life, savings would not be known until after the calculation of a shipment's cost, a complicated and time-consuming procedure.

Table III.3: Savings From Using Direct Air Shipment Instead of Air Freight Forwarder Classified by Air Freight Forwarder Shipping Costs of More Than $100[a]

Stratum	Shipment	Savings
$100-$499	1	$32
	2	62
	3	190
	4	140
	5	96
	6	99
	7	78
	8	130
	9	66
	10	75
	11	48
	12	160
	13	110
	14	145
	15	159
	16	200
	17	109
	18	100
	19	153
	20	127
	21	45
	22	90
	23	157
	24	92
	25	155
	26	167
	27	125
	28	59
	29	78
	30	78

(continued)

Stratum	Shipment	Savings
	31	154
	32	158
	33	199
	34	96
	35	83
	36	105
	37	61
	38	138
	39	142
	40	170
	41	108
	42	113
	43	139
	44	121
	45	143
	46	147
	47	232
	48	192
	49	182
	50	182
	51	71
	52	98
	53	63
	54	132
	55	65
	56	57
	57	128
	58	140
	59	141
	60	113

(continued)

Stratum	Shipment	Savings
	61	149
	62	201
	63	112
	64	188
	65	164
	66	94
	67	127
	68	156
	69	64
	70	198
	71	121
	72	208
	73	0
	74	134
	75	121
$500 or more	1	$431
	2	500
	3	502
	4	320
	5	259
	6	457
	7	304
	8	276
	9	404
	10	270
	11	255
	12	373
	13	252
	14	348
	15	336

(continued)

Stratum	Shipment	Savings
	16	264
	17	321
	18	360
	19	375
	20	251
	21	285
	22	210
	23	445
	24	288
	25	462

[a]In real life, savings would not be known until after the calculation of a shipment's cost, a complicated and time-consuming procedure.

We will assume that the evaluators decide to take a preliminary sample of 50 items to estimate the total savings and the sampling error and to determine the final sample size. In the illustrations of the three allocation methods, we will use the SRO-STATS computer package printouts and their options.

Suppose the evaluators decide to select a preliminary sample of 30 items with 15 items from the stratum less than $100, 10 items from the stratum $100 to $499, and 5 items from the stratum $500 or more. Using the GAO-approved random number generator, table III.4 provides the random numbers selected and the value of the savings for the selected item.

Table III.4: Preliminary Sample Selection

Stratum	Shipment	Savings[a]
Less than $100	95	$34
	70	27
	40	26
	7	23
	143	25
	87	34
	147	23
	65	41
	60	31
	125	33
	122	33
	134	10
	5	11
	109	45
	62	19
$100 to $499	57	$128
	11	48
	14	145
	67	127
	5	96
	40	170
	56	57
	29	78
	50	182
	15	159
$500 or more	23	$445
	21	285
	7	304
	16	264
	15	336

[a]In real life, savings would not be known until after the calculation of the shipment's cost, a complicated and time-consuming procedure.

From this preliminary sample, we can estimate that the average savings per shipment was $84.98 and that we are 95-percent confident that the true but unknown mean is in the range of $74.68 to $95.28.

Proportional Allocation

Assume that the evaluators decide to use proportional allocation and that the sample size that is decided on is 50 shipments. The SRO-STATS program provides the following information:

1. With a total sample of 50 shipments, the evaluators should take 30 shipments from the stratum less than $100 (they have already taken 15, so that they need to take an additional 15 shipments), 15 shipments from the stratum $100 to $499 (10 already taken and, thus, they need to take 5 additional shipments), and 5 shipments from the stratum $500 or more.

2. Assuming that the standard deviations for each stratum remain fairly constant, the precision for the estimated mean will decline from $10.30 to $8.71.

If we select the additional sample items and calculate the savings for each shipment selected, we can show in table III.5 the sample results. The results in figure III.1 were calculated from these sample results with the SRO-STATS program.

Table III.5: Calculated Savings

Stratum	Shipment	Savings
Less than $100	95	$34
	70	27
	40	26
	7	23
	143	25
	87	34
	147	23

(continued)

Stratum	Shipment	Savings
	65	41
	60	31
	125	33
	122	33
	134	10
	5	11
	109	45
	62	19
	71	29
	19	39
	55	22
	11	23
	129	50
	37	35
	84	42
	41	13
	33	15
	14	24
	75	31
	22	19
	121	36
	26	16
	13	19
$100 to $499	57	$128
	11	48
	14	145
	67	127
	5	96
	40	170
	56	57
	29	78
	50	182
	15	159

(continued)

Stratum	Shipment	Savings
	44	121
	68	156
	16	200
	69	64
	19	153
$500 or more	23	$445
	16	264
	21	285
	15	336
	7	304

Figure III.1: Results From SRO-STATS Calculation

FOR CONFIDENCE LEVEL = 95%

STRATUM	UNIVERSE SIZE	SAMPLE SIZE	MEAN	PRECISION (OF MEAN)	ESTIMATE	PRECISION (OF TOTAL)
<100	150	30	27.60	3.28	4140	492
100-499	75	15	125.60	21.49	9420	1612
500+	25	5	326.80	55.81	8170	1395
TOTAL	250	50	86.92	8.75	21730	2188

Neyman Allocation

For Neyman allocation, we used the preliminary sample of 30 sample shipments and the SRO-STATS program to allocate the final sample of 50 items into the three strata. The results of that program showed the following results:

1. The final sample size should be allocated in the following manner. The stratum less than $100 should have a sample size of 11 items (we have already selected 15, so the final sample size will be 54), the stratum $100 to $499 should have a sample size of 26 (we have already selected 10, so we need to select an

additional 16 items), and the stratum $500 or more should have a sample size of 13 (we have already selected 5, so we need to select an additional 8 items).

2. Precision will decrease from the preliminary sample of $10.30 to $6.15 under the assumption that the standard deviations in the three strata remained fairly constant.

Using the GAO-approved random number generator, we selected the additional sample shipments to reach our final sample size. Table III.6 shows the final sample items.

Table III.6: Final Sample Items Under Neyman Allocation

Stratum	Shipment	Savings
Less than $100	95	$34
	40	26
	143	25
	147	23
	60	31
	122	33
	5	11
	62	19
	70	27
	7	23
	87	34
	65	41
	125	33
	134	10
	109	45
$100 to $499	57	$128
	14	145
	5	96
	56	57
	50	182

(continued)

Stratum	Shipment	Savings
	44	121
	16	200
	19	153
	12	160
	42	113
	26	167
	61	149
	39	142
	11	48
	67	127
	40	170
	29	78
	15	159
	68	156
	69	64
	24	92
	49	182
	59	141
	46	147
	54	132
	53	63
$500 or more	23	$445
	7	304
	15	336
	3	502
	12	373
	18	360
	5	259
	21	285
	16	264
	22	210

(continued)

Stratum	Shipment	Savings
	10	270
	17	321
	1	431

If we enter these results into the SRO-STATS program for stratified sampling, the estimated results from the program show that we estimate that the average savings per shipment was $89.05 and that we are 95-percent confident that the true but unknown mean is in the range $83.27 to $94.83 (89.05 plus or minus 5.78). Notice that this precision is somewhat less than we estimated it would be from the preliminary sample but remember that the final sample size was 54, not 50.

Judgmental Allocation

To illustrate judgmental allocation, we can assume that the evaluators decide to compute savings for all shipments of $500 or more, for 15 sample shipments from the $100 to $499 stratum, and for 10 sample shipments from the stratum of less than $100. The sample results are shown in table III.7.

Table III.7: Sample Results Under Judgmental Allocation

Stratum	Shipment	Savings
Less than $100	35	$10
	57	33
	62	19
	80	36
	81	21
	113	30
	117	37
	126	23
	129	50
	135	33

(continued)

Stratum	Shipment	Savings
$100 to $499	2	$62
	5	96
	10	75
	20	127
	23	157
	28	59
	30	78
	31	154
	37	61
	38	138
	43	139
	45	143
	61	149
	66	94
	67	127
$500 or more	1	$431
	2	500
	3	502
	4	320
	5	259
	6	457
	7	304
	8	276
	9	404
	10	270
	11	255
	12	373
	13	252
	14	348
	15	336
	16	264
	17	321
	18	360

(continued)

Stratum	Shipment	Savings
	19	375
	20	251
	21	285
	22	210
	23	445
	24	288
	25	462

Using the SRO-STATS program, we estimate that the savings per shipment would be $84.89, and we are 95-percent confident that the true but unknown mean savings is between $78.45 and $91.33 (the precision of the mean is $6.44).

A Comparison of the Sampling Errors

To summarize, our sampling errors are $8.75, $5.78, and $6.44 for proportional, Neyman, and judgmental allocation, respectively. As can be seen, Neyman allocation gave the smallest precision. Judgmental allocation provided the next best estimate of the precision, because the stratum with the largest variation was sampled 100 percent. However, a comparison with Neyman allocation reveals that the sample in the top stratum (shipments of $500 or more) was too large and that the sample from the middle stratum was too small. Proportional allocation gave the largest precision because half the sample was drawn from the stratum with the least variation. Judgmental allocation does not always give a better result than proportional allocation. If 60 percent of the sample had been allocated to the bottom stratum (shipments less than $100), the precision would have been greater than that obtained with proportional allocation.

It is interesting to see what would have happened if a simple random sample of 50 shipments had been taken from the entire population without regard to stratification. In the random sample of 50 that was drawn to illustrate this point, the precision was $25.73, which is roughly 2.94 times the result obtained with proportional allocation. In fact, using the SRO-STATS program, we estimate that a sample size of 212 would be needed to obtain the same precision as the Neyman allocation.

Stratified Sampling for Attributes

Stratified sampling can also be used when sampling for attributes. Compared with simple random sampling, stratified sampling may slightly reduce the precision. It also allows the development of separate estimates for individual strata, if this is necessary, provided that the sample sizes in the strata are sufficiently large. However, the disadvantage of stratifying when sampling for attributes is that the increased (smaller) precision is usually not worth the additional work that is required.

To illustrate stratified sampling for attributes, we can assume that the evaluators are reviewing civilian payroll records at three military bases. Since the records are separately maintained at each location, the population is stratified by location. The evaluators decide to select independent random samples of 100 payroll records at each base (judgmental allocation), with the results shown in table III.8.

Table III.8: Independent Random Samples of Payroll Records

Location	Population	Sample	With errors
1	1,100	100	45
2	1,500	100	5
3	400	100	20
Total	**3,000**	**300**	**70**

Using the stratified estimation program from the SRO-STATS program, we can estimate that the error rate in the population of civilian employees at the three military bases was 21.667 percent, and we are 95-percent confident that the true but unknown error is 21.667 percent plus or minus 4.107 percent. We can also estimate that the number of civilian employees whose paychecks were in error was 650 and that we are 95-percent confident that the true but unknown number of employees is in the range of 527 to 773.

Assume that the evaluators would like to know what sample sizes would be required, using proportional allocation and Neyman allocation, to reduce the precision of the stratified percentage to 2 percent at the 95-percent confidence level. Using the proportional allocation option of the SRO-STATS program, we can estimate the sample size needed to reduce the precision to 2 percent under the assumption that the percentage findings would not vary greatly from the original sample. To accomplish the reduction of the precision to 2 percent requires the evaluators to select the following sample sizes: location 1, 334; location 2, 455; and location 3, 121 (total = 910).

Likewise, we can use the Neyman allocation option of the stratified program to estimate the required sample size to reduce the precision to 2 percent. The required sample size was location 1, 421; location 2, 253; and location 3, 123 (total = 797). Thus, with Neyman allocation, the specified precision can be obtained by using 113 fewer payroll records than with proportional allocation.

Other Topics on Stratification

Stratification has some other advantages not mentioned above. One is that, by careful stratification, the evaluators can maximize the dollars protected; that is, they can review the maximum dollar amounts of transactions, documents, or accounts with a given sample size. Another advantage is that, by careful stratification, the evaluators can maximize the number of errors discovered and corrected; that is, they can include error-prone items in one stratum and relatively error-free items in another. They can then sample more heavily from the error-prone stratum. Last, but not least, stratification permits the development of estimates for the individual strata, if such estimates are needed.

Some Practicalities of Stratification

A word should be said about the realities of stratification in most applications. Textbook illustrations usually assume that (1) the strata were designed by the sampler to reduce the amount of the precision, (2) the stratification is based on the variable being estimated, and (3) the standard deviations are known or can be computed for the variable being estimated. In real life, however, the strata are often defined by the objectives of the job or the physical location or arrangement of the population.

The standard deviations for the variable being estimated are not known and must be computed from preliminary samples taken in each stratum. The stratum boundaries and sample sizes (both overall and within strata) must be calculated from some variable other than the variable being estimated, because this is the only characteristic available. For example, consider the direct air shipment versus air freight forwarder problem. The strata boundaries were based on air freight forwarder shipping costs, but in real life, shipping costs, not savings, might have to be used to compute standard deviations for the sample size calculations. The basis for this is the belief that the variance of the characteristic being estimated is highly correlated with the variance of the variable used to set the strata boundaries.

Sometimes, from a practical point of view, it is just as efficient to use general rules of thumb to determine the total sample size and to allocate the sample to the individual strata on the proportion of the individual stratum total to the grand total. For example, assume that the air freight forwarder shipping costs were as shown in table III.9.

Table III.9: Assumed Air Freight Forwarder Costs

Stratum	Shipping costs	Percent of total
Less than $100	$13,200	24
$100-$499	25,300	46
$500 or more	16,500	30
Total	**$55,000**	**100**

If we were allocating a sample of 50 items on this basis, we would draw 12 items from the first stratum (24 percent of 50), 23 items from the second (46 percent of 50), and 15 items from the third (30 percent of 50).

This allocation method assumes that stratum standard deviations of the savings are roughly proportional to the stratum means of the shipping costs. In practice, this method often works out fairly close to the results obtained by using Neyman allocation. If the results are not as precise as required, it may be necessary to increase the sample size in one or more of the strata.

If the air freight forwarder shipping costs were not known for each stratum, another possibility would be to assume that the mean shipping cost per stratum equals the stratum midpoint. (The stratum of $500 or more presents a problem because it is open ended; however, we can often make a reasonable assumption about midpoints for open-ended strata.) Here, for example, assume midpoints of $50 for stratum 1, $300 for stratum 2, and $600 for stratum 3. The allocation of a sample of 50 to the various strata is shown in table III.10.

Table III.10: Allocation of a Sample of 50

Stratum	Population	Midpoint	Product	Percent
Less than $100	150	50	7,500	16
$100 to $499	75	300	22,500	50
$500 or more	25	600	15,000	34
Total	**250**		**45,000**	**100**

Thus, when we allocate our sample of 50 items to the individual strata, we have a sample of 8 items for stratum 1 (16 percent of 50), 25 items from stratum 2 (50 percent of 50), and 17 items from stratum 3 (34 percent of 50).

As can be seen, this method gives a different allocation of the sample from that obtained by using the Neyman allocation, but it is very close and far better than the results obtained by using proportional allocation. The advantage of this method is that it does not require prior information about the characteristic being estimated.

Guidelines on Constructing Strata

A word about the construction of strata is in order. Evaluators may well ask: How many strata should we have? Where should we set the strata boundaries? A body of mathematical theory has been developed on how to determine the optimum number of strata and how to set the boundaries, but a discussion of this theory is beyond the scope of this paper. However, some general rules of thumb can be given.

In sampling for variables, when the evaluators frequently use stratification to make the precision smaller, six strata are usually sufficient. If the number of strata is increased beyond six, the change in the precision is usually not worth the extra work required.

As for setting the strata boundaries, if the population is listed in ascending or descending order of value, the boundary locations usually become obvious when the evaluators scan the list. If the population is so large that it is not possible to list every item, the evaluators may list a sample of items, say 5 or 10 percent, sorted in order of value, to set the boundaries. Another possibility is to base the boundaries on a frequency distribution of the items. For example, consider the frequency distribution on transaction dollar amounts in table III.11.

Table III.11: Frequency Distribution on Transaction Dollar Amounts

Amount of transaction	Number of transactions	Dollar amount	Cumulative amount
Less than $10	4,063	$30,879	$30,879
$10-$19	3,323	61,190	92,069
$20-$29	3,063	70,151	162,220
$30-$39	2,544	95,909	258,129
$40-$49	1,424	67,926	326,055
$50-$59	839	46,145	372,200
$60-$69	593	39,434	411,634
$70-$79	397	30,768	442,402
$80-$89	352	30,976	473,378
$90-$99	274	26,770	500,148
$100-$199	194	34,338	534,486
$200-$299	183	47,214	581,700
$300-$399	119	39,746	621,446
$400-$499	61	27,023	648,469
$500-$599	41	22,427	670,896
$600-$699	29	18,879	689,775
$700-$799	20	15,080	704,855
$800-$899	23	19,040	723,895
$900-$999	9	8,757	732,652
$1,000 or more	10	14,800	747,452
Total	**17,561**	**747,452**	

After examining the frequency distribution, the evaluators may decide to set the boundaries at less than $10, $10 to $19, $20 to $49, $50 to $99, $100 to $199, $200 to $499, $500 to $999, and $1,000 or more. This gives a set of strata in which the upper stratum boundary is about twice the lower, except in the lowest and highest strata. Another possibility is to divide the overall total by the required number of strata to obtain the average dollar amount per stratum. Then the boundaries are set where the cumulative totals are closest to the product of the stratum numbers and the average dollar amount per stratum.

For example, suppose we wanted to have six strata. We divide the total dollar amount, $747,452, by 6 to obtain $124,575, or the average dollar amount per stratum. We then multiply this amount by each of the stratum numbers to obtain the results in table III.12.

Table III.12: Dollar Amount Multiplied by Strata Numbers

Stratum number	Product
1	124,575
2	249,150
3	373,725
4	498,300
5	622,875
6	747,450

Then we look at the cumulative amounts column in the frequency distribution, locate the amounts that are closest to the products, and set the stratum boundaries there. Using this system, we obtain the boundaries in table III.13.

Table III.13: Strata Boundaries

Cumulative amount	Stratum boundary
$92,069	Less than $20
258,129	$20-$39
372,200	$40-$59
500,148	$60-$99
621,446	$100-$399
747,452	$400 or more

This method makes the total dollar amount in all strata approximately equal. If equal sample sizes are allocated to all strata, the result approximates Neyman allocation, provided that the standard deviations of the variables being estimated are fairly proportional to the stratum means of the variable being used to set the boundaries.

Cluster Sampling

Here is an example of a two-stage cluster sample problem in which the clusters or primary sampling units are selected by simple random sampling.

While reviewing the procurement activity of a government-operated scientific laboratory, the evaluators decide to determine the dollar amount of prompt payment discounts that were lost on invoices paid during the past fiscal year, either because the invoices were not paid promptly or because the discount was not taken. The invoices paid during the fiscal year, together with their supporting documentation, are tied in 2,100 bundles, containing varying quantities of invoices. Thus, each bundle can be defined as a cluster.

Setting the confidence level at 95 percent, the evaluators decide to take a preliminary random sample of 40 clusters. The invoices on which discounts were offered can be identified easily by examining the terms of sale. However, calculating the actual amount of discounts lost involves (1) determining how long the discount period was and whether the invoice was paid within the discount period and (2) multiplying the invoice amount by the discount rate (percent) if the invoice was not paid within the discount period or was paid promptly but the discount was not taken.

The evaluators decide that if a sample bundle contains less than 10 invoices on which discounts were offered, they will calculate the discount lost for all such invoices. But to save work, if the bundle contains 10 or more invoices on which discounts were offered, they will calculate the discount lost for a random sample of 5 invoices. Random number sampling can be used to select the sample invoices. The results are shown in table III.14.

Table III.14: Random Number Sample

	Requisition		
Sample bundle	Number in bundle	Number sampled	Amounts of discounts lost
1	7	7	44,0,32,17,0,0,0
2	60	5	0,0,0,37,46
3	7	7	0,0,0,0,0,50,6
4	48	5	18,0,46,0,32
5	68	5	25,22,0,0,0
6	65	5	2,0,0,0,35
7	70	5	0,0,39,0,0
8	55	5	0,30,38,15,0
9	12	5	19,0,0,0,19
10	4	4	23,0,0,0
11	38	5	29,25,0,7,0
12	9	9	0,3,1,0,37,9,0,0,0
13	12	5	4,32,0,0,0
14	70	5	14,0,0,10,0
15	8	8	0,0,30,0,20,0,0,0
16	48	5	18,0,7,0,28
17	42	5	0 0,0,0,0,21
18	2	2	38,6
19	5	5	0,0,47,0,0
20	3	3	31,0,43
21	65	5	10,5,5,0,18
22	5	5	15,35,44,0,0
23	15	5	0,15,0,13,41

(continued)

Sample bundle	Requisition: Number in bundle	Requisition: Number sampled	Amounts of discounts lost
24	7	7	50,0,37,3,33,0,0
25	6	6	42,10,17,41,0,0
26	24	5	8,15,1,0,0
27	2	2	42,42
28	8	8	3,17,18,29,0,38,0,0
29	9	9	0,1,21,13,0,0,43,0,0
30	5	5	22,0,0,39,50
31	36	5	0,0,0,32,0
32	3	3	42,0,0
33	4	4	0,0,0,11
34	15	5	0,16,22,0,46
35	13	5	21,38,0,0,0
36	72	5	0,0,29,0,0
37	33	5	0,8,0,0,0
38	85	5	42,0,7,15,0
39	5	5	23,0,16,24,0
40	18	5	18,0,0,0,41

Using the SRO-STATS program, we would estimate that the laboratory lost during the past fiscal year in prompt pay discounts $558,600 with a precision at the 95-percent confidence level of $190,859.70. In other words, the true but unknown total amount of discounts lost would be in the interval between $367,740.30 and $749,459.70.

Another method could have been used to estimate the dollar amount of discounts lost. This is the ratio-to-size estimate, which uses the technique of ratio estimation discussed in chapter 5. To use this method, we would have to know the total number of invoices paid during the fiscal year.

In this paper, we have not gone into mathematical methods for computing sample sizes with cluster sampling; however, a few general comments are in order. According to a rule of thumb, we should try to have at least 30 clusters in the sample. A sample with as few as 20 clusters may sometimes give fairly precise results. However, for various reasons that need not be discussed here, estimates developed from a cluster sample are usually much less precise than estimates developed from a simple random sample consisting of the same number of items.

If we are using two-stage cluster sampling and we can make a choice between sampling more items within the cluster or reducing the number of items sampled within the cluster and increasing the number of clusters in the sample, it is better to increase the number of clusters and reduce the number of items sampled within the cluster. This will practically always yield more precise estimates.

Dollar Unit Sampling

The selection of any one particular method of sample design requires that the evaluators have determined (1) their objectives and (2) the characteristics of the population from which the sample is to be drawn. Dollar unit sampling is designed to allow the evaluators to make a statement about the amount of error (both overstatements and understatements) in the population of interest. Dollar unit sampling is designed to enable a conclusion similar to the following to be drawn: "From the results of our sample, we are 95-percent confident that the amount of dollar error in the audited population does not exceed $xxxxxx (where the xxxxxx is the estimated population total based on the sample results)." The evaluators then compare the value of $xxxxxx with some measure of materiality that enables them to reach a conclusion about the acceptability of the reported book value of the population of interest.

However, before using dollar unit sampling, the evaluators must have information about the population of interest to determine if the assumptions used in this sampling method are met. The two basic assumptions are

1. The error rate in the population should be small (less than 10 percent) and the population should contain 2,000 or more items. (The use of the Poisson probability for evaluation of the sample requires this feature.)

2. The amount of error in any item in the population cannot be more than the reported book value of the item. That is, if the book value of an item is $100, the amount of error in the balance cannot exceed $100.

If the assumptions are valid for the population of interest and the conclusion as stated above coincides with the objectives of the job, the evaluators should consider using dollar unit sampling.

Dollar unit sampling is a modified form of sampling for attributes that permits dollar conclusions about the total dollar amount of error in the population. Unlike the simple random approach of normal attribute sampling, it focuses on the individual dollars. For example, suppose we are evaluating a population of inventory parts that has 5,000 stock numbers and a book value of $1,000,000. Instead of viewing the population as 5,000 stock items from which to select a sample, we would think of the population as made up of 1,000,000 individual dollar units from which we would draw our sample. However, when an individual dollar unit is selected, it acts as a hook and brings into the sample the entire balance for that stock item containing the individual dollar.

Computer Software Packages

This appendix briefly describes two comprehensive statistical sampling and analysis packages, SAS and SPSS, as well as several other packages. SAS and SPSS are popular in GAO. They have been available to GAO for many years and are now accessible to microcomputer users as well as mainframe computer users. Mainframe access is provided by large time-sharing computer facilities such as the National Institute of Health's Division of Computer Research and Technology and the Public Health Service's Parklawn Computer Center. Microcomputer access for GAO users is provided through GAO's contractual arrangements with software vendors.

This appendix should not be misconstrued to mean that GAO endorses SAS and SPSS. Literally hundreds of statistical software packages are on the market. Some packages cover a broad range of statistical procedures; others concentrate on specific statistical techniques and specialized research. Some packages work on a wide variety of computers and operating systems; others are available on only a few types of computers. Some packages include data retrieval, graphics, and reporting capabilities; others concentrate on statistical procedures. Most packages are easy to use, comprehensively tested, well documented, and powerful tools for selecting samples and calculating statistical results.

Statistical calculations are not limited to statistical programs. Many GAO users adapt retrieval packages such as DYL-AUDIT and spreadsheet packages such as Lotus 123 to select samples and perform simple calculations. Since these packages are not primarily intended for statistical use they must be used with extreme caution.

All statistical software packages can easily be misapplied and their results easily misinterpreted. Sometimes this is the result of the user's failure to completely understand the assumptions the package makes in performing the analysis. In addition, the packages are periodically updated in a way that modifies statistical procedures and corrects errors. Whenever you decide to use any software for statistical sampling or statistical analysis, it is highly recommended that assistance be requested from the appropriate technical assistance group.

SAS

SAS is a comprehensive statistical language. It includes a variety of statistical, quality assurance, operations research, graphics, mapping, and data retrieval procedures. SAS also includes a matrix manipulation language to write customized statistical routines and a macro language to facilitate repetitive statistical computations.

SAS offers the ultimate in power. What is lacking in preprogrammed procedure can be supplemented by user-written procedures. Its data manipulation procedures are the most powerful available, including complex data transformations. The package is consistent between all versions—microcomputer, minicomputer, and mainframe—so that the user need not be concerned about the actual computer processing of the data.

SAS is easy to use. All versions provide interactive, fill-in-the-blank programming in a "windowed" environment as well as traditional programming. SAS allows users to read data files from other packages, like SPSS, and allows SAS programs and data files to use features from other programs such as BMDP. Specialized packages, such as SUDAAN, allow SAS users to calculate the results of complex sampling designs.

SAS has several shortcomings. The link between user-written data formats and SAS data files is sloppy. SAS cannot directly read a Lotus 123 file and requires that the user convert the Lotus 123 file to a DIF file, an inconvenience. SAS is not for the occasional user. SAS documentation is spread out in half a dozen or more large manuals that can be intimidating.

GAO has licenses for microcomputer and minicomputer versions that provide access to the software. The mainframe version is available through the National Institute of Health Computer Center, the Parklawn Computer Center, and other time-sharing systems.

Documentation is available directly from the SAS Institute. Registered users of the National Institute of Health Computer Center can obtain copies of the documentation through its PUBWARE system.

SPSS

SPSS is a comprehensive statistical language, including a variety of statistical procedures. It includes statistical, quality assurance, graphics, mapping, and data retrieval procedures.

SPSS is easy to learn. To help the new user, the microcomputer version provides a context-sensitive help window with menu selection programming. It is also easy to find information in the manuals once a specific procedure has been identified. SPSS data manipulation procedures are intuitive. SPSS's microcomputer graphics easily produce standard business graphics through the Harvard Graphics package. SPSS allows users to read data files from other packages such as SAS, Lotus 123, and BMDP.

SPSS has several shortcomings. The mainframe and microcomputer versions are not totally consistent. Character data manipulation is limited. The menu system of writing programs is easy to learn but not easy to use. SPSS does not compute weighted standard deviations.

GAO has licenses for microcomputer and minicomputer versions that provide access to the software. Users may access the mainframe version of SPSS through the National Institute of Health Computer Center and other time-sharing systems.

Documentation is available directly from SPSS, Inc. Registered users of the National Institute of Health Computer Center can obtain copies through the PUBWARE system.

Gauss-A Matrix Calculator Program

Gauss is a program that performs matrix calculations without introducing programming issues. It is written for use on microcomputers and is suitable for teaching and statistical applications. Statistical analyses such as regression, correlation, principal components, canonical correlations, and discriminant analysis are easy with the matrix formulations.

STATPRO

STATPRO is another microcomputer statistical package. It lets users do almost everything they do on a mainframe on the microcomputer, including descriptive statistics, regression, ANOVA, and factor and cluster analysis, just to name a few. Its awesome power is not limited to number crunching. Users can plot all their results in four color graphics, such as scatter diagrams, triangle and regression plots, histograms, and pie charts. It also has data base management capabilities that make entering, manipulating, transforming, and editing data quick and easy.

BMDP

The BMDP computer programs are designed to aid data analysis by providing methods ranging from simple data display and description to advanced statistical techniques. Data are usually analyzed by an iterative "examine and modify" series of steps. The data are first examined for unreasonable values, graphically and numerically. If unreasonable values are found, they are checked and, if possible, corrected. An analysis is then performed. This analysis may identify other inconsistent observations or indicate that further analyses are needed. The BMDP programs are designed to handle all steps in an analysis from the simple to the sophisticated.

DYL-AUDIT

DYL-AUDIT, or DYL, which runs only on IBM computers, is primarily a data retrieval and report package. Some of GAO's technical assistance groups use it. It is oriented more toward use by data processing specialists than by data analysis specialists. The documentation is reasonably well written and fairly clear. The language that the user writes in is very close to English and includes terminology commonly used by IBM data processors. The case selection part of the language allows taking many kinds of samples by using a few lines of programming.

The manuals contain useful descriptions of the sampling procedures and explain why options should be chosen. This documentation can be useful even as background information in sampling. The self-documenting features are somewhat limited and the data modification features are limited. DYL does have the ability to do grouped frequency counts. It also computes many kinds of subtotals but does not compute all marginal subtotals. It is very useful for extracting data already on an IBM computer. It is not typically used by GAO for calculating the estimates to the population based on the results of the sample. The output is very clearly organized and includes worksheets that aid in data gathering. The presentation of results is very flexible, and it is relatively easy to prepare the data so that they can be moved to other machines or programs.

DYL-AUDIT is recommended for extracting data when there is a very large amount of data and machine efficiency is a major consideration. Because jobs requiring the use of this package are very large, a specialist who knows DYL should be consulted.

IMSL

IMSL, International Mathematical and Statistical Library, contains computational subroutines written in the computer language FORTRAN and has been tested by mathematical and statistical computation specialists. The writers of the library adhere to rigorous standards for computation and documentation. The written documentation is clearly organized and always gives detailed instructions on the input and output for the subroutines. The routines usually contain checks for many kinds of errors. IMSL is oriented toward high-level specialists who need to create programs for functions not included in the standard packages. Versions are available for many sizes and brands of computers. The manuals describe the procedures but do not explain why options should be chosen.

IMSL is recommended when new programs must be written. Typically, it is used by statisticians who can write in FORTRAN.

SRO-STATS

SRO-STATS is a series of nine statistical programs that GAO's Seattle office developed to serve as a training and audit tool on GAO's microcomputers. These programs have been reviewed and the formulas approved by GAO's statistical staff. SRO-STATS includes programs for

1. simple random variable sampling;
2. simple random attribute sampling;

3. simple random sampling for making ratio, regression, and difference estimation when both the "book" and audited values are known;

4. making estimates for stratified sampling for both variables and attributes;

5. making estimates for two-stage cluster sampling for both variables and attributes;

6. testing whether the difference between two means from simple random samples is statistically significant;

7. testing whether the difference between two proportions from simple random samples is statistically significant;

8. testing whether the difference between three or more means from simple random samples is statistically significant;

9. testing whether the difference between two or more proportions from simple random samples is statistically significant.

Their advantage over most sampling programs are that they are very easy to use and they provide estimated sample sizes or precision for a variety of user-specified conditions. Once the main menu is called up, the user need only select the program and then respond to a series of questions. The programs are merely tools for performing statistical calculations; however, someone with a statistical background should still supervise the sampling procedures and the interpretation of results.

Bibliography

Aggarwal, Y. P. Better Sampling: Concepts, Techniques and Evaluation. New York: Apt Books, 1988.

Anderson, T. W. An Introduction to Multivariate Statistical Analysis, 2nd ed. New York: John Wiley and Sons, 1984.

Arkin, H. Handbook of Sampling for Auditing and Accounting, 3rd ed. New York: McGraw-Hill, 1984.

Bailey, A. D. Statistical Auditing. New York: Harcourt Brace Jovanovich, 1981.

Blalock, H. M. Social Statistics, 2nd ed. New York: McGraw-Hill, 1979.

Box, G. E. P., T. Leonard, and C. Wu. Scientific Inference, Data Analysis, and Robustness: Proceedings of a Conference Conducted by the Mathematics Research Center of The University of Wisconsin—Madison, November 4-6, 1981. New York: Academic Press, 1983.

Cochran, W. G. Planning and Analysis of Observational Studies (edited by L. E. Moses and F. Mosteller). New York: John Wiley and Sons, 1983.

Cochran, W. G. Sampling Techniques, 3rd ed. New York: John Wiley and Sons, 1977.

Deming, W. E. Sample Design in Business Research. New York: John Wiley and Sons, 1990.

Deming, W. E. Some Theory of Sampling. New York: Dover Publications, 1966.

Dixon, W. J., and F. J. Massey. Introduction to Statistical Analysis, 4th ed. New York: McGraw-Hill, 1983.

Guy, D. M. An Introduction to Statistical Sampling in Auditing. New York: John Wiley and Sons, 1981.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow. Sample Survey Methods and Theory, vol. 1. New York: John Wiley and Sons, 1953.

Huntsberger, D. V., and P. Billingsley. Elements of Statistical Inference, 4th ed. Boston: Allyn and Bacon, 1977.

Ijiri, Yuji, and Robert S. Kaplan. "The Four Objectives of Sampling in Auditing: Representative, Corrective, Protective, and Preventive." Management Accounting, 52:60 (December 1970), 42-44.

Jessen, R. T. Statistical Survey Techniques. New York: John Wiley and Sons, 1978.

Johnson, N. L., and H. Smith (eds.). New Developments in Survey Sampling. A Symposium on the Foundations of Survey Sampling held at the University of North Carolina, Chapel Hill, North Carolina. New York: Wiley-Interscience, 1969.

Kalton, G. Introduction to Survey Sampling. Beverly Hills, Calif.: Sage Publications, 1983.

Kish, L. Statistical Design for Research. New York: John Wiley and Sons, 1987.

Kish, L. Survey Sampling. New York: John Wiley and Sons, 1965.

Kruskal, W., and F. Mosteller. "Representative Sampling." International Statistical Review, 47 (1979), 13-24, 111-27, 245-65; 48 (1980), 169-95.

Kuebler, R. R., and H. Smith. Statistics: A Beginning. New York: John Wiley and Sons, 1976.

Levy, P. S., and S. Lemeshow. Sampling for Health Professionals. Belmont, Calif.: Wadsworth, 1980.

Mendenhall, W. Introduction to Probability and Statistics, 6th ed. Boston: Duxbury Press, 1983.

Moore, D. S., and G. P. McCabe. Introduction to the Practice of Statistics. New York: W. H. Freeman and Company, 1989.

Ott, L. Introduction to Statistical Methods and Data Analysis, 3rd ed. Boston: Duxbury Press, 1988.

Raj, D. Sampling Theory. New York: McGraw-Hill, 1968.

Roberts, D. Statistical Auditing. New York: American Institute of Certified Accountants, 1978.

Rowntree, D. Statistics Without Tears: A Primer for Nonmathematicians. New York: Charles Scribner's Sons, 1981.

Scheaffer, R. L., W. Mendenhall, and L. Ott. Elementary Survey Sampling. Belmont, Calif.: Wadsworth, 1979.

Sincich, T. Business Statistics by Example, 2nd ed. San Francisco: Dellen Publishing Company, 1986.

Slakter, M. J. Statistical Inference for Educational Researchers. Reading, Mass.: Addison-Wesley Publishing Company, 1972.

Slonim, M. J. Sampling in a Nutshell, rev. ed. New York: Simon and Schuster, 1960.

Smith, G. Statistical Reasoning. Boston: Allyn and Bacon, 1988.

Snedecor, G. W., and W. G. Cochran. Statistical Methods, 8th ed. Ames, Iowa: The Iowa State University Press, 1989.

Sudman, S. Applied Sampling. New York: Academic Press, 1976.

Sukhatme, P. V., and B. V. Sukhatme. Sampling Theory of Surveys With Applications, 2nd ed. Ames, Iowa: Iowa State University Press, 1970.

Tennant-Smith, J. BASIC Statistics. Boston: Butterworth, 1985.

Williams, B. A Sampler on Sampling. New York: John Wiley and Sons, 1977.

Glossary

Array
An arrangement of a series of items according to the values of the items, usually from largest to smallest or smallest to largest.

Attribute
As used in attribute sampling, an inherent quality or characteristic that an item either has or does not have. It can be either a simple quality or characteristic, such as being or not being a high school graduate, or a complex one, such as strongly agree or not strongly agree (made up of the responses strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree).

Attribute Sampling
In attribute sampling, the selected sampling units are measured or evaluated in terms of whether they have the attribute of interest, and some statistical measure (statistic) is computed from these measurements to estimate the proportion of the population that has the attribute.

Bias
The existence of a factor that causes an estimate made on the basis of a sample to differ systematically from the population parameter being estimated. Bias may originate from poor sample design, deficiencies in carrying out the sampling process, or an inherent characteristic of the measuring or estimating technique used.

Census
A complete enumeration of a population. This is a 100-percent sample.

Central Limit Theorem	In its simplest form, the theorem states that for sample data from a population with a finite variance, the sampling distribution of the sample means approaches the normal distribution as the sample size becomes larger and larger. This theorem is of fundamental importance in probability and statistics, as it justifies the application of normal distribution theory to a great variety of statistical problems.
Cluster Sample	A simple random sample in which each sampling unit is a collection of elements.
Confidence Coefficient	A measure (usually expressed as a percentage) of the degree of assurance that the estimate obtained from a sample differs from the population parameter being estimated by less than the measure of precision (sampling error).
Confidence Interval	A range of values that is believed, with a preassigned degree of confidence, to include the particular value of some parameter or characteristic being estimated. The degree of confidence is related to the probability of obtaining by random samples ranges that are correct See also Sampling Error.
Confidence Level	See Confidence Coefficient.
Consistent Estimate	An estimate that tends to be closer to the true but unknown value of the parameter as the size of the sample increases.

Correlation	The interdependence between two sets of numbers; a relation between two quantities, such that when one changes, the other changes. Simultaneous increasing or decreasing is called "positive correlation"; one increasing and the other decreasing is called "negative correlation."
Data	The results of an experiment, census, survey, and any kind of process or operation.
Degrees of Freedom	A random sample of size n is said to have n - 1 degrees of freedom for estimating the population variance, in the sense that there are n - 1 independent deviations from the sample mean on which to base such an estimate.
Descriptive Statistics	Although this term has been used to refer only to tabular and graphic presentations of statistical data, nowadays it is used more broadly to refer to any treatment of data that does not involve generalizations.
Deviation	The difference between the particular number and the average of the set of numbers under consideration.
Dispersion	The extent to which the elements of a sample or the elements of a population are not all alike in the measured characteristic, are spread out, or vary from one another.

Domains of Interest	Classes into which a population may be subdivided so that separate estimates can be developed for each domain. This is different from stratification, because a domain of interest can extend across several strata and because the classification may be based on the sample data.
Estimate.	See Statistic.
Exhaustive Sampling	The 100-percent inspection of a population. See Census.
Finite Population Correction Factor	Abbreviated FPC, a multiplier that makes adjustments for the sampling efficiency gained when sampling is without replacement and when the sample size is large (greater than 5 or 10 percent) with respect to the population size. This multiplier reduces the sampling error for a given sample size or reduces the required sample size for a specified measure of precision (in this case, desired sampling error).
Frequency Distribution	A table in which data are grouped into classes and the number of items that fall into each class is recorded.
Geometric Mean	The geometric mean of n positive numbers is the positive nth root of their product.
Histogram	A graphic representation of a frequency distribution.

Inference	The making of statements, or the process of drawing judgments about a population on the basis of random samples, in such a manner that the probability of making correct inferences is determinable under various alternative hypotheses about the population being sampled.
Interquartile Range	The distance between the first and third quartiles of a distribution. Covers the middle half of the values in the frequency distribution.
Interval Estimation	The estimation of a parameter in terms of an interval, called an "interval estimate," for which one can assert with a given probability (or degree of confidence) that it contains the actual value of the parameter. See <u>Confidence Interval</u>.
Judgment Sample	Unlike a probability sample, a sample in whose selection personal judgment plays a significant part. Though judgment samples are sometimes required by practical considerations, and may lead to satisfactory results, they do not lend themselves to analysis by standard statistical methods.
Kurtosis	The relative peakedness or flatness of a distribution. A distribution that is more peaked and has relatively wider tails than the normal distribution is said to be "leptokurtic." A distribution that is less peaked and has relatively narrower tails than the normal distribution is said to be "platykurtic."

Mean	The sum of all the values in a set of observations (this can be either a sample or a census) divided by the number of observations. Also known as "average" or "arithmetic mean," it indicates the typical value for a set of observations.
Median	The middle measurement when the items are arranged in order of size or, if there is no middle one, then the average of the two middle ones. If five students make the grades 15, 75, 80, 95, and 100, the median is 80.
Mode	The most frequent value of a set of numbers. If more students (of a given group) make 75 than any other one grade, then 75 is the mode.
Monte Carlo Method	Any procedure that involves statistical sampling techniques in obtaining a probabilistic approximation to the solution of a mathematical or physical problem.
Optimum Allocation	A method of allocating a sample to strata by taking into account not only the difference in strata population sizes and standard deviations but also the differences in the costs of collecting data for the various strata.
Parameter	A measure such as mean, median, standard deviation, or proportion that is calculated or defined by using every item in the population.
Percentile	The value that divides the range of a set of data into two parts such that a given percentage of the measures lies below this value.

Population	All the members of a group to be studied as defined by the evaluators; the total collection of individuals or items from which a sample is selected.
Precision	See Sampling Error.
Probability	The ratio of the number of outcomes that will produce a specific event to the total number of possible outcomes, or the likelihood that specific events will occur, expressed as a proportion or percentage.
Probability Sampling	The selection of a sample by some random method to obtain information or draw conclusions about a population. All possible samples, and thus each item in the population, have a known and specified (nonzero) probability of being drawn.
Proportional Allocation	In stratified sampling, the allocation of portions of the total sample to the individual strata so that the sizes of these subsamples are proportional to the sizes of the corresponding strata. For instance, if a stratified sample of 100 students is to be taken from among the 400 freshmen, 300 sophomores, 200 juniors, and 100 seniors attending an undergraduate school, proportional allocation requires 40, 30, 20, and 10 students be chosen from these four classes.
Quartile	The 25th, 50th, and 75th percentiles are the first, second, and third quartiles.

Quartile Deviation

A measure of variation, also called the "semi-interquartile range," that is given by half the difference between the first and third quartiles; hence, the average amount by which the first and third quartiles differ from the median.

Random Decimal Digits

A table of digits 0 through 9 arranged so that digits may be randomly selected according to any procedure, subject to the sole restriction that a digit's selection be influenced only by its location in the table. Its purpose is to permit the drawing of random samples.

Random Number Sampling

A sampling method in which combinations of random digits, within the range of the number of items in a population, are selected by using one of the random number generation methods until a given sample size is obtained. For example, if a sample of 60 items is required from a population numbered 1 through 2,000, then 60 random numbers between 1 and 2,000 are selected.

Random Selection

A selection method that uses an acceptable method of generating random numbers in a standard manner. The method minimizes the influence of nonchance factors in selecting the sample items.

Ratio Estimate

An estimate of a population parameter that is obtained by multiplying the known population total for another variable by a ratio of appropriate sample values of the two variables. See also Regression Estimate.

Regression	The line of average relationship between the dependent (or primary) variable and the independent (or auxiliary) variable.
Regression Coefficient	A measure of change in a primary variable associated with a unit change in the auxiliary variable.
Regression Estimate	An estimate of a population parameter for one variable that is obtained by substituting the known total for another variable into a regression equation calculated on the basis of sample values of the two variables. Note that ratio estimates are special kinds of regression estimates.
Sample	A portion of a population that is examined or tested in order to obtain information or draw conclusions about the entire population.
Sampling Error	Each estimate generated from a probability sample has a measurable precision, or sampling error, that may be expressed as a plus or minus figure. A sampling error indicates how closely we can reproduce from a sample the results that we would obtain if we were to take a complete count of the population using the same measurement methods. By adding the sampling error to and subtracting it from the estimate, we can develop upper and lower bounds for each estimate. This range is called a "confidence interval." Sampling errors and confidence intervals are stated at a certain confidence level. For example, a confidence interval at the 95-percent confidence level means that in 95 of 100 instances, the sampling procedure we used would produce a confidence interval containing the population value we are estimating.

Sampling for Attributes	See Attribute Sampling.
Sampling for Variables	See Variable Sampling.
Sampling Frame	A means of access to a population, usually a list of the sampling units contained in the population. The list may be printed on paper, a magnetic tape file, a file of punch cards, a computer disk, or a physical file of such things as payroll records or accounts receivable.
Sampling Units	The elements into which a population is divided; they must cover the whole population and not overlap, in the sense that each element in the population belongs to one and only one unit.
Sampling With Replacement	A sampling method in which each item selected for a sample is returned to the population and can be selected again. In this method, the population can be regarded as infinite.
Sampling Without Replacement	A sampling method in which an item selected for a sample is "used up": it is not returned to the population and cannot be selected again. In this method, the population can be regarded as finite.
Scientific Sampling	See Probability Sampling.

Simple Random Sample	A sample obtained by a selection of items from the population is a simple random sample if each item in the population has an equal chance of being drawn.
Standard Deviation	A numerical measurement of the dispersion, or scatter, of a group of values about their mean. Also called "root mean square" deviation.
Standard Error	The standard deviation of the sampling distribution of a sample statistic.
Statistic	A measure, such as a mean, proportion, or standard deviation, derived from a sample and used as a basis for estimating the population parameter.
Statistical Estimate	A numerical value assigned to a population parameter on the basis of evidence from a sample.
Statistical Sampling	See Probability Sampling.
Statistics	Methods of obtaining and analyzing quantitative data. The following aspects are applicable only in reference to some phase of the experimental logic of quantitatively measured, variable, multiple phenomena: (1) inference from samples to populations by means of probability (commonly called "statistical inference"); (2) characterizing and summarizing a given set of data without direct reference to inference (called "descriptive statistics"); (3) methods of obtaining samples for statistical inference (called "sampling statistics").

Strata	Two or more mutually exclusive subdivisions of a population defined in such a way that each sampling unit can belong to only one subdivision or stratum.
Stratified Random Sample	If the population to be sampled is first subclassified into several subpopulations called "strata," the sample may be drawn by taking random samples from each stratum. The samples need not be proportional to the strata sizes.
Systematic Selection With a Random Start	A sampling method in which a given sample size is divided into the population size in order to obtain a sampling interval. A random starting point between 1 and the sampling interval is obtained. This item is selected first; then every item whose number or location is equal to the previously selected item plus the sampling interval is selected, until the population is used up.
Tolerable Error	The specified precision or the maximum sampling error that will still permit the results to be useful.
True Mean	The mean of a population; the term is meant to emphasize the distinction between a sample mean and the constant (though unknown) mean of a population.
Universe	See Population.
Variable	As used in variable sampling, a characteristic having values that can be expressed numerically or quantitatively and that may vary from one observation to another. Examples are the dollar amount of error in a voucher, a quantity shipped, and the height of a person.

Variable Sampling

In variable sampling, the selected sampling units are measured or evaluated (in terms of dollars, pounds, days, and so on), and some statistical measure (statistic) is computed from these measurements to estimate the population parameter or measure.

Papers in This Series

This is a flexible series continually being added to and updated. The interested reader should inquire about the possibility of additional papers in the series.

The Evaluation Synthesis. Transfer paper 10.1.2.

Content Analysis: A Methodology for Structuring and Analyzing Written Material. Transfer paper 10.1.3, formerly methodology transfer paper 3.

Designing Evaluations. Transfer paper 10.1.4.

Using Structured Interviewing Techniques. Transfer paper 10.1.5.

Using Statistical Sampling. Transfer paper 10.1.6.

Developing and Using Questionnaires. Transfer paper 10.1.7, formerly methodology transfer paper 7.

Case Study Evaluations. Transfer paper 10.1.9.

Prospective Evaluation Methods: The Prospective Evaluation Synthesis. Transfer paper 10.1.10.

Quantitative Data Analysis: An Introduction. Transfer paper 10.1.11.

Ordering Information

The first copy of each GAO report is free. Additional copies are $2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

U.S. General Accounting Office
P.O. Box 6015
Gaithersburg, MD 20877

Orders may also be placed by calling (202) 275-6241.

United States
General Accounting Office
Washington D.C. 20548

Official Business
Penalty for Private Use $300

First Class Mail
Postage & Fees Paid
GAO
Permit No. G100