United States
Department
of Agriculture

Forest Service

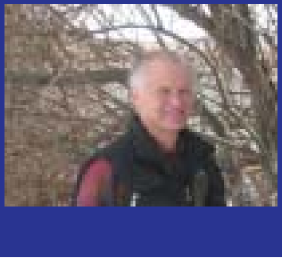
**Rocky Mountain
Research Station**

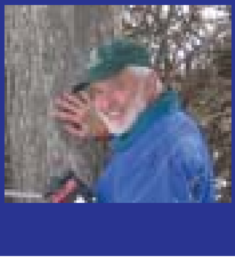
General Technical
Report RMRS-GTR-126

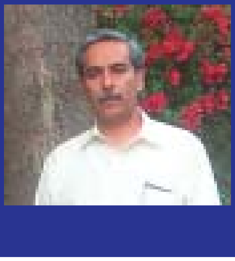
April 2004

Statistical Techniques for Sampling and Monitoring Natural Resources

Hans T. Schreuder, Richard Ernst, and Hugo Ramirez-Maldonado







CORRECTIONS

Page 26: In Equation (31), a n_h should be added to the denominator, i.e.,

$$v(\bar{y}_{st}) = \sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{(N_h - n_h)}{N_h \times n_h} s_h^2$$

Page 27: The fourth equation from the top and the line following it should be

$$v(\bar{y}_{st}) = \frac{1}{10^2}\left(5^2 \frac{5-2}{5\times 2} 2 + 5^2 \frac{5-2}{5\times 2} 50\right) = \frac{780}{200} = 3.9$$

Then $\hat{Y}_{st} = 10 \times 8.5 = 85$ with variance estimate $v\left(\hat{Y}_{st}\right) = 100 \times 3.90 = 390$

Schreuder, Hans T.; Ernst, Richard; Ramirez-Maldonado, Hugo. 2004. **Statistical techniques for sampling and monitoring natural resources.** Gen. Tech. Rep. RMRS-GTR-126. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 111 p.

Abstract

We present the statistical theory of inventory and monitoring from a probabilistic point of view. We start with the basics and show the interrelationships between designs and estimators illustrating the methods with a small artificial population as well as with a mapped realistic population. For such applications, useful open source software is given in Appendix 4. Various sources of ancillary information are described and applications of the sampling strategies are discussed. Classical and bootstrap variance estimators are discussed also. Numerous problems with solutions are given, often based on the experiences of the authors. Key additional references are cited as needed or desired.

Acknowledgments

We owe a considerable measure of gratitude to reviewers for valuable comments. Chistopher Kleinn, Steen Magnusson, Ray Czaplewski, Rudy King, and Keith Rennolls reviewed the entire book, Geoff Wood revised an earlier draft, and Jule Caylor took the time to check the remote sensing section and added some updates to it. Mike Williams reviewed the coarse woody debris section, Jim Baldwin and Paul Geissler the wildlife section, Ron McRoberts and Nick Crookston the modeling section, Frank Roesch and Paul Patterson the section on discrete variable sampling, Charles Scott the section on multilevel sampling, and Gretchen Moisen the section on small area estimation. Gary Boyak reviewed and partially rewrote the sections on GIS and GPS, Henry Lachowski reviewed the section on GIS, and Jeff Goebel reviewed the section on sampling for rare events. Tim Gregoire and Kim Iles made useful suggestions on how to handle boundary trees. Lane Eskew did an excellent job of reviewing the manuscript from an editorial point of view and making sure that it conformed with acceptable standards of publication. We owe a considerable debt of gratitude to Ray Czaplewski for suggesting updating Frank Freese's book and facilitating the subsequent writing of this book.

The Authors

Hans T. Schreuder is a Mathematical Statistician (retired) with the USDA Forest Service's Rocky Mountain Research Station in Fort Collins, CO. **Richard Ernst** is a Mensurationist with the USDA Forest Service's Forest Management Service Center (Washington Office) in Fort Collins, CO. **Hugo Ramirez-Maldonado** is a Director General with the National Institute on Forestry, Agriculture and Animal Husbandry Research in Mexico City, Mexico.

You may order additional copies of this publication by sending your mailing information in label form through one of the following media. Please specify the publication title and series number.

Fort Collins Service Center

Telephone	(970) 498-1392
FAX	(970) 498-1396
E-mail	rschneider@fs.fed.us
Web site	http://www.fs.fed.us/rm
Mailing address	Publications Distribution Rocky Mountain Research Station 240 West Prospect Road Fort Collins, CO 80526

Rocky Mountain Research Station
Natural Resources Research Center
2150 Centre Avenue, Building A
Fort Collins, CO 80526

Contents

I. Introduction

The purpose of this book is to serve as a complete introduction to the statistical techniques of sampling natural resources starting at a very basic level and progressing to more advanced methods. We describe supplementary tools and materials and identify key references for readers wishing to pursue the subject further. Considerable material is based on direct experiences of the authors. We include introductory material, much of which is taken from the excellent introductory book of Freese (1962). These sections in Freese's book are expressed in a compelling and still relevant manner. A good example is Chapter V: Sampling Methods for Discrete Variables. More advanced readers can skip these sections. To facilitate reading in general, we dispense with the proof of properties of estimators, such as their unbiasedness and how the variances of estimators are derived. Schreuder spent most of his career with the USDA Forest Service working for Forest Inventory and Analysis Program (FIA). Ernst teaches numerous short courses in forest inventory and sampling and provides consultation on such methods to inventory and measurement specialists in the National Forest System (NFS). Ramirez-Maldonado has considerable experience in teaching courses in forest sampling, inventory, and modeling as well as consulting for Mexican agencies in forest inventory and monitoring.

There are several good introductory books available on sampling. The book by Johnson (2000) is very basic and gives extensive information. It is dated, however, in that it does not cover more recent advances in the field. The books by Freese (1962) and deVries (1986) are still useful in providing several of the basic concepts, the latter going further afield in what is available. Freese's book has the additional advantage that it is available in Spanish. Shiver and Borders (1996) provide a modernized version of Freese (1962) material with some emphasis on typical forestry methods.

More advanced books are available, too, in forestry. To a large degree this book represents a simplification of the book by Schreuder and others (1993). The book by Iles (2003) reveals why he is such a good speaker and writer; it is a delight to read and is worth examining for practical suggestions. Gregoire and Valentine (2004), as judged from the outline of their book, appear to have considerable overlap with ours, but it is more likely to be tightly written and aimed at a more sophisticated audience. It is also more limited in its objectives but helpfully contains numerous proofs showing the properties of various estimators. Arvanitis and Reich (2004) provide the most complete description of geostatistical methods in forestry, methods which rely heavily on models at this time. For readers interested in obtaining a full understanding of how and why probabilistic sampling methods work, the classical books Sarndal and others (1992) and Cassel and others (1977) are mandatory readings and surprisingly easy to read given their strong theoretical orientation. The book by Cochran (1977) is still quite popular with many practitioners and presents the basic sampling theory well (with a few exceptions). It too is available in Spanish.

II. Objectives of Sampling and Monitoring

Before discussing the methodology of survey sampling, some brief comments about statistics are desirable. What is it? Generally, statistics should be thought of as "systematized common sense." It protects us from jumping to conclusions. A good example is the classical experiment on the effect of aspirin on headaches. Initial tests showed that it helped 80 percent of the people who tried it, certainly a phenomenal success rate. Then someone had the idea of trying a placebo. It showed a 60 percent success rate, indicating that although aspirin was useful, many people apparently did not need it to relieve their headache. Because so many things are open to different interpretations and because the USA is such a litigious society, statistics have become critical in many fields of endeavor including natural resources. Hence statistically valid sampling methods and surveys have become important in generating reliable and legally defensible estimates. Surveys, also called inventories, are the basis for strategic, management, and project planning by generating a reliable data base. Since a census or complete count of resources would be prohibitively costly and time consuming, sampling of forest resources began around the beginning of the 20th century (Schreuder and others 1993).

Before designing a sample survey, the objectives must be clearly defined. Many surveys are started with a single limited objective, e.g., we just want to know how much timber is available for harvesting from a certain area or what areas may support numerous (often unspecified) rare plant species. Many of these surveys are subsequently used for other purposes. Often the novice sampler spends much money collecting data on a large number of items and then cannot answer specific questions. If a survey is planned, particularly a large-scale one, it is highly desirable to critically examine the data to be collected to be sure that the survey truly addresses the requirements of the proposed users. Questions to be asked may be: Are objectives well defined and attainable? Are measurements on weeds really needed? If tree quality is considered an important variable but cannot be measured reliably, is it still worth measuring? Remember: you may be blamed for failure to plan ahead even though your users may have assured you that they had only limited objectives or that timber really was not more important than other information or that they did not have enough money to finance the survey properly.

Why Sample?

The purpose of sampling is to draw inferences about a population of interest such as what is the average height of trees in a forest. The overall field of inference is very broad and technical and is discussed in more detail in the section on inference in Appendix 1. There are many ways of making inferences and people can differ vehemently on how to get the necessary information and how to draw inferences/conclusions on the basis of that information. We focus on a limited part of the field of inference, the drawing of probabilistic samples from finite populations, and the inferences typically made with such data.

Most decisions in life are made with incomplete knowledge. Your physician may diagnose disease from a few drops of blood or microscopic sections of tissue; a consumer judges watermelons by the sound they emit when thumped; and we select toothpaste, insurance, vacation spots, mates, and careers with but a fragment of the total information necessary or desirable for complete understanding. Our hope is that the drops of blood or the tissue samples represent the non-sampled portions of the body, the sounds of the watermelons indicate the maturity of the melon, and that the advertising claims present an honest representation of the truth.

Partial knowledge is normal. The complete census is rare; the sample is more usual. A ranger advertises timber sales with estimated volume, estimated grade yield and value, estimated cost, and estimated risk. Bidders take the accuracy and reliability of this information at their own risk and judgment. The nursery specialist sows seed whose germination is estimated from a tiny fraction of the seed lot, and at harvest estimates the seedling crop with sample counts in the nursery beds. Managers plan the maintenance of recreation areas based on past use and experience**.**

Typically we collect information from a population. We call this a sample. We then summarize this information in some manner. Probably the most widely used sample summarization is the sample mean. Assume that we can take samples of 3 units from some population, then our judgment often is based on the mean of the three, i.e., $\bar{y} = \frac{y_1 + y_2 + y_3}{3}$ with y_i the value for the variable on sample unit i, i = 1,2,3.

However desirable a complete census may seem, there are good reasons why sampling is often preferred. In the first place, complete measurement or enumeration may be impossible, e.g., determining the exact amount of wood in a forest would cost many times its value; the nursery specialist would be better informed if the germination capacity of all the seed to be sown was known, but the destructive nature of the germination test precludes testing every seed. Clearly where testing is destructive, some sort of sampling is inescapable. Use of a recreation area is not known until the season is over; judging what resources are needed to manage an area has to be based on previous information.

Sampling frequently provides the essential information at a far lower cost than complete enumeration. For large populations especially, the data collected is often more reliable. There are several reasons why this might be true. With fewer observations to be made and more time available, crews will become less tired and remain more committed to careful measurement. In addition, a portion of the savings in cost could be used to buy better instruments and to employ or train highly qualified personnel. Certainly careful measurements on 5 percent of the units in a population could provide more reliable information than careless measurements on 100 percent of the units. Finally, sample data can usually be collected and processed in a fraction of the time required for a complete inventory, so that the information obtained is timely.

Planning Your Survey

Objectives

The first step is to define the objectives, also considering the possibility that they may be amplified, modified, or extended over time. Successful surveys often are continued over time with additional objectives added on. For large-scale forest surveys such as the one in the USA, where a national survey is conducted by the Forest Inventory and Analysis (FIA) staff of the FS, the objectives have changed over time. This is what one should expect with successful surveys. The objectives of most surveys are covered by the following set of objectives for large-scale surveys such as FIA:

1. Generate current status estimates such as acreage in forest area, amount of wood volume by species groups, mortality, timber volume available for harvest, etc.
2. Monitor change in the above and other parameters over time.
3. Establish procedures required for identifying possible cause/effect relationship hypotheses.
4. Establish procedures required to prove or document cause-effect. Since cause-effect can rarely be established with survey data and usually requires followup experimentation, it is important to indicate what can and cannot be done in this regard.
5. Provide in-place information for managers by proper development of such techniques as using maps in conjunction with small area estimation.
6. Provide timely information for decision makers.
7. Maintain a reliable database with comprehensive documentation and reliable archiving, and encourage better and more analyses.

Originally FIA was established for objective #1. Over time as concern for timber supplies became more critical, #2 became as important. #3 became important in the 1980s with the controversy of apparent declining forest growth in the state of Georgia. #4 almost always requires both survey sampling and experimentation. Much research is being done on #5 right now. #6 will be a critical one for management areas, and #7 has always been important but will become even more so with an annualized inventory where industry and the states can and want to analyze the data much more frequently and independently. FIA has changed from a periodic to an annual approach, so instead of collecting data every 10 years, between 10 and 20 percent of the national plots will be measured every year with reports on a state basis every 5 years. This was done to meet the increasing need for more current information.

A classic on planning surveys is Hush (1971). This provides useful supplementary reading to help in such planning. Particularly, his appendix "Sample outline for preparing inventory plans" would be a very useful starting point for people contemplating a brand new survey.

Information to be collected

For most objectives, existing probabilistic survey designs can be used or modified readily to satisfy one's objectives. Cause-effect is a different issue, dealt with in more detail later. Often the credibility of the results from an inventory and monitoring system is of paramount importance. This requires stringent criteria in the survey. Some or all of the following criteria and thoughts should generally be considered for any reasonable survey (Schreuder and Czaplewski 1992):

1. Only variables with negligible measurement errors or ones that can be efficiently calibrated with variables with such negligible measurement errors should be used. Subjective observations have high rates of measurement error and unpredictable biases that can compromise their utility; objective measurements can readily be justified usually even if more expensive to collect.
2. Destructive sampling cannot be allowed on permanent sample plots. Off-plot destructive sampling might be acceptable in the immediate vicinity of the plot.
3. Exact locations of permanent sample plots need to be kept secret to avoid biased treatment by landowners or visitors or excess visitations that damage vegetation or soil and make it unrepresentative of the population.
4. Define all variables so no confusion is possible.
5. Define some variables as truth being measured from remote sensing sources rather than by ground sampling. Remote sensing interpretation can be more efficient and accurate than field measurement for some variables, preventing the inadvertent disturbance of plots by field crews and denying access to plots by landowners. In some cases there is some flexibility in the definition of a variable of interest; for example, canopy cover measured from low-altitude photos as opposed to estimated from ground samples.
6. Don't protect sample plots differently from the remainder of the population as is often done for growth and yield plots.

Closely related to this is the importance of defining variables with the following characteristics:

1. Those that can be accurately measured on aerial photos so that field sampling is not necessary. For example, in some cases this may be possible with percent canopy cover or change in area of mature forest but not in change in commercially suitable forest.
2. Variables that can easily be measured in the field such as tree mortality and number of trees. Such variables may often be correlated with variables measured on aerial photos.
3. Variables difficult or expensive to measure in the field. Examples are tree volume, tree crown condition, and horizontal and vertical vegetation profiles. Surrogates should be sought such as basal area for volume.
4. Variables for which a within-growing season distribution may be desired such as rainfall amounts and distribution, ozone concentrations, chemical composition of tree components, and symptoms of arthropod and microbial effects on trees. This requires more than one visit in a season, something we often cannot afford in surveys.
5. Variables for which destructive sampling is required such as soil and needle samples for chemical composition and tree cores for tree growth and dendrochronological studies. How this may affect remeasurement over time needs to be considered carefully.

The following design objectives are critical:

1. Collect data on explanatory/stress variables such as rainfall deficiency, low soil moisture, exposure to pollution, etc. This type of data usually cannot be collected on plots yet but are essential in building reliable models.
2. Simplicity in design. This provides flexibility over time and ease in analysis.
3. Consistency of design over time. This simplifies change estimation and identifying possible cause-effect hypotheses.
4. Flexibility to address new environmental or management issues while maintaining design consistency.
5. Flexibility to incorporate new measurement technologies while maintaining design consistency.
6. Ability to treat each sample unit as a population. This is important, for example, in classifying each sample to estimate acreage in forest types. This means, for example, no missing data for a sample unit because of the design used.
7. Use interpenetrating sampling or similar methods so sampling intensity can be readily increased in time and space if needed. This is a nice feature of annualized inventories if handled properly.
8. Provide flexibility to accommodate replacement of plots to deal with damage caused by the measurement process (for example, trampling or destructive sampling) or denial of access to plots by private landowners—for example sampling with partial replacement.
9. Ability to handle missing data such as plots being inaccessible or landowners denying access (as noted by C. Kleinn, inaccessibility may also be caused by land mines or wildlife such as elephants and lions). Inaccessibility is best handled by setting aside a separate stratum for such plots and clearly stating the estimated size of that stratum and how estimates if any are generated for it. Implement a strong quality assurance program so that true changes in sample plots over time will not be confounded with changes in measurement error or subtle details in measurement protocol.
10. Consider use of several plot designs at the same sample locations. Although this complicates data collection, it may well be required when a large suite of parameters is of interest. For example, for number of trees and total basal area of trees, very different plot designs are efficient (fixed area and variable radius plots respectively).

Developing the sampling approach

Given the objectives of one’s survey, the idea is to develop the most cost efficient approach to reach those objectives. Most of the remainder of the book is devoted to designing such efficient sampling approaches by giving the reader insight into methods available and how and when to use them. Basically what we are developing are sampling strategies, which consist of how to collect the data—what we call the design and how to use the data to estimate the quantity of interest, i.e., the estimation process.

III. Sampling Concepts and Methodologies

Sampling Frame

We all make inferences about populations based on what is typically a biased sample. Knowing this often drives us crazy in talking with people. For example: Person A: teenagers are terrible drivers! Person B: Oh, on what do you base that statement? Person A: Well, when I was driving last week I got cut off twice by teenagers.

A sampling frame is a complete list of the sample units that can potentially be selected in the population. To avoid biased inferences such as the one above about the teenagers, make sure that the population defined for sampling is the one of interest as well as the sample units it consists of. For example, suppose we are interested in the following two parameters about the city of Colima in the state of Colima, Mexico, and Fort Collins, Colorado, USA:

1. The average income of each household.
2. The average area of land owned by landowners.

In these examples, a household would initially be the possible sample unit in the first and a landowner in the second. How do we proceed to list the two populations of interests? Is this important? It is critical that each sample unit in the population has a positive probability of being selected for the sample and that we know what that probability is. Using a list of people with phones is clearly not a complete list of all people in either city, but it is certainly less complete in Colima. Since a list of households may not be available for either city, different sample units may be considered such as city blocks for which there generally would be a list (how to make inferences about households when the sample unit is a city block will be explained later under cluster sampling). Obtaining a list of all landowners would probably be fairly easy for both cities since all landowners presumably pay taxes and hence can be found on a tax listing for the cities.

Selecting a representative sample from a population is easiest when we have a complete list of all units (sample units) from which to draw a representative sample. For example, assume N_1 = number of ha and N_2 = number of trees in the same forest. Clearly we often know N_1 but rarely N_2. If all N_1 ha are listed, we can take a simple random sample (SRS) of n_1 ha plots. Then we have a random sample of plots but with a different number of trees per plot usually. It is generally not easy (and inefficient) to draw a random sample of trees from a population of trees.

It is possible to draw a random sample without having a sampling frame. Then a list of units is available only after sampling (see Sect 3.4 p.72-73 in Schreuder and others 1993, where a procedure, described by Pinkham 1987 and Chao 1982, is discussed). However, the procedure is awkward to implement.

Purposive and Representative Sampling

In purpose sampling, also called non-probabilistic or model-based sampling, samples are selected more or less deliberately. This can be done on the basis of the judgment of the sampler of what is a desirable sample or whatever sample happens to be convenient to collect. This is generally not considered a representative sample of the population of interest.

The idea of selecting a representative sample from a population was extensively discussed in the literature dating from early in the 20th century (Johnson and Kotz 1988, Vol. 8, p. 77-81). Eight methods of selection have been described, but the method of random or probability sampling discussed below is generally favored. The basic idea is to select a sample completely by chance selection to ensure that there is no personal bias involved in selecting it. To do this, we use randomization in selecting the sample, i.e., select a sample from a deliberately haphazard arrangement of observations. To implement this, we use probabilistic sampling, which involves sampling in such a way that:

1. Each unit in the population is potentially selected with a known positive probability of selection.
2. Any two units in the population have a joint positive probability of selection.

Problem: A property comprises 100,000 ha of forest, range, and water and the owner wishes to find out what is there. Develop a sampling method that satisfies the two conditions above.

Answer: There are numerous ways of doing so. One approach: divide a map of the property in 100,000 1-ha plots and randomly select 20 of these 1-ha plots for classification into the categories forest, range, and water. This satisfies the two conditions. Estimation may be difficult because some of the plots may contain more than one of the classes; but how to deal with that will be covered in the estimation theory later on.

A sampling strategy is comprised of a sampling design and associated estimation theory where a sampling design states formally how samples are selected. Potential sample units can have equal or unequal probabilities and joint probabilities of selection meeting the above two criteria. This flexibility leads us to the designs discussed later, i.e., SRS, stratified sampling, cluster sampling, and variable probability sampling.

To complete the picture of our sampling strategy, we need estimators accordant with the design selected. Sampling designs with the simplest estimator, additional estimators, and some general sampling procedures are discussed below.

Populations, Parameters, Estimators, and Estimates

The central notion in any sampling problem is the existence of a population. It is helpful to think of a population as a collection of units with values of interest attached. The units are selected in some way and the values of interest are obtained from the selected ones in some manner, either by measurement or observation. For example, we may imagine a 40-ha tract of timber in which the unit being observed is the individual tree and the value being observed is tree height. The population is the collection of trees with heights on the tract. The aggregate number of branches on these same trees would be another population as would the number of trees with hollows suitable for animal nesting.

To characterize the population as a whole, we often use certain constants of interest that are called parameters, usually symbolized with Greek letters. Some examples are the mean number of trees per plot in a population of plots; the proportion of living seedlings in a pine plantation; the total number of shrub species in a population; and the variability among the unit values.

The objective of sample surveys is usually to estimate such parameters. In the past, we tended to estimate the population mean or total of one or more variables. Nowadays, we are often also interested in possible explanations of why a parameter is a certain value. The value of the parameter as estimated from a sample will hereafter be referred to as the sample estimate or simply the estimate, symbolized by Roman letters. The mathematical formula generally used to generate an estimate is called an estimator.

Whenever possible, matters will be simplified if the units in which the population is defined are the same as those to be selected in the sample. If we wish to estimate the total weight of earthworms in the top 15 cm of soil for some area, it would be best to think of a population made up of blocks of soil of some specified dimension with the weight of earthworms in the block being the unit value. Such units are easily selected for inclusion in the sample, and projection of sample data to the entire population is relatively simple. If we think of individual earthworms as the units, selection of the sample and expansion from the sample to the population may be very difficult if not impossible.

Problem: How would you go about sampling an ant nest to estimate the number of ants in it?

Answer: If the nest can be destroyed, one can scoop it up, take samples of a certain volume from it at random, and count the number of ants in each of these samples. If it cannot be destroyed there is really no obvious way to take a representative sample from the nest to count the ants.

Bias, Accuracy, and Precision

When estimating population parameters, one wishes to obtain good estimates close to the true values at a reasonable cost. When only a part of a population is measured, some estimates may be high, some low, some close, and some far from the true value. An estimate that is close to the true value is considered accurate. If the person selecting or measuring a sample is prejudiced in some way, then the estimate may be biased. For example, if one was interested in the recreational preferences of visitors to a park and interviewed a sample of 99 women and 1 man, one might feel uneasy about bias in the results, just as one might if the results for a survey of 50 men and 50 women showed a heavy preference for fishing in a park not noted for fishing and also knowing that the interviewer was an avid fisherman.

Though most people have a general notion as to the meaning of bias, accuracy, and precision, statisticians have well-defined expressions for them because they are crucial in their area of expertise, as follows:

Bias—Bias is a systematic distortion that can arise when selecting a sample, during its measurement, or when estimating the population parameters.

Bias due to sampling selection arises when certain units are given a greater or lesser representation in the sample than in the population. This is not compensated for in estimation. Assume for example that we are estimating the recreational preferences of visitors to a park and we only interview people on weekends. The results will be biased because weekday users had no opportunity to appear in the sample.

Measurement bias can result. For example, if seedling heights are measured with a ruler from which the first half-cm is missing, all measurements will be one-half cm too large and the estimate of mean seedling height will be biased. In studies involving tree counts, some observers may always include a tree that is on the plot boundary while others may consistently exclude it. Both routines are sources of measurement bias. In timber cruising, the volume table selected or the manner in which it is used may result in bias. For example, a volume table constructed from data of tall trees will give biased results when used without adjustment on short trees. Similarly, if a cruiser consistently overestimates tree heights, volume tables using heights as input will be biased. The only practical way to minimize measurement bias is by meticulously training the crews in measurement procedures and the use, care, and calibration of instruments.

The technique used to estimate the population parameters from the sample data is also a possible source of bias. For example, if the recreation preference on two national forests is estimated by taking a simple arithmetic average of the preferences recorded on each forest, the result may be seriously biased if the area of one forest is 500,000 ha and it has a million visitors annually and the other is 100,000 ha in size with only 10,000 visitors annually. A better overall estimate would be obtained by weighing the estimates for the two forests in proportion to their sizes and/or their number of visitors. Another example of this type of bias occurs in the common forestry practice of estimating the average diameter of trees in a forest from the diameter of the tree of mean basal area. The latter procedure actually gives the square root of the mean squared diameter, which is not the same as the arithmetic mean diameter unless all trees are exactly the same size.

Selection and measurement biases are rarely acceptable, particularly if the data are of interest to several users. However, estimation bias can often be acceptable since some biased estimators are much better than unbiased ones, the bias being often trivial and the results more precise than those achieved using the unbiased procedures. Acceptable biased estimators are usually asymptotically unbiased estimators, defined as follows:

Asymptotically unbiased—If the bias of an estimator approaches 0 as the sample size approaches the population size, the estimator is considered to be asymptotically unbiased. Such an estimator used to be called consistent, for example in Cochran (1977).

Precision and accuracy—A biased estimate may be precise but it can never be accurate. Among statisticians *accuracy* refers to the success of estimating the true value of a quantity; *precision* refers to the extent of clustering of sample values about their own average, which, if biased, cannot be the true value.

A target shooter who puts all of his shots in the inner circle of a target might be considered both accurate and precise; his friend who puts all of her shots in the outer circle and near the 12 o'clock position would be considered equally precise but nowhere near as accurate; another friend who always hits the target at some random location would be unbiased but neither accurate nor precise. This is illustrated in Figure 1 below.

A forester making a series of careful diameter measurements at a fixed position on the bole of a tree with a caliper, one arm of which is not at right angle to the graduated beam, would achieve precise but not accurate results. Since the caliper is not properly adjusted, the measured values will be off the true value (bias) and the diameter estimate will be inaccurate. If the caliper was properly adjusted but used carelessly, the measurements would be unbiased but neither accurate nor precise.

Generally we strive to use estimators that predict a parameter more reliably than competing estimators where reliability is usually measured by the ratio of the mean square errors of the estimators. Such estimators are called efficient estimators.

Variables: Continuous and Discrete

Variation is a fact of life. Coping with some of the sampling problems created by variation is an important part of making valid inferences. All objects have characteristics such as size, shape, and color. A characteristic that varies from unit to unit is called a variable. In a population of trees, tree height is a variable, as are tree diameter, number of cones, volume, form class, and species. The number of people in each recreation group is a variable, as are their sex, their age, etc.

A variable that is expressed in a numerical scale of measurement, any interval of which may, if desired, be subdivided into an infinite number of values, is said to be continuous, e.g., time recreating, height, weight, precipitation, and volume. Qualitative variables and those that are represented by integral values or ratios of integral values are said to be discrete. Two forms of discrete data may be recognized: attributes and counts. An attribute refers to units classified as having or not having some specific quality; examples of attributes might be species or whether trees are alive or dead. Results are often expressed as a proportion or percent, e.g., incidence of rust in slash pine seedlings, survival of planted seedlings, and the percentage of users from a particular country of a recreation area. A count refers to units described by a number, such as number of people in a recreation group, number of weevils in a cone, and number of sprouts on a tree stump.

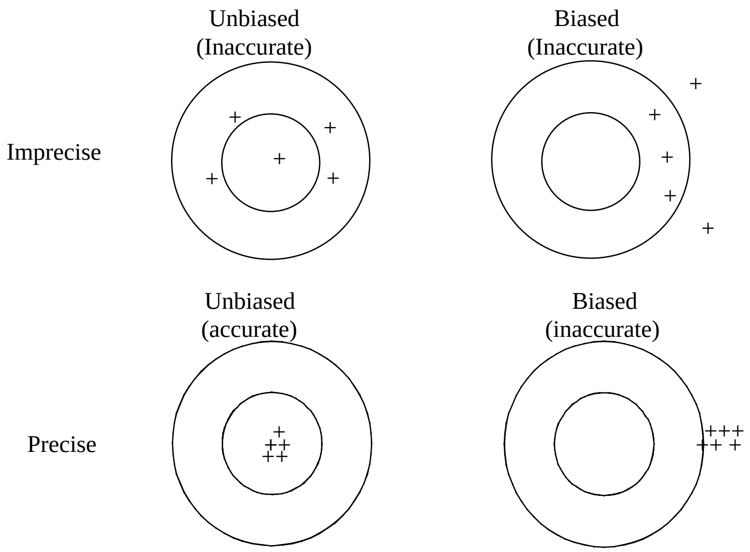


Figure 1. An example of bias, precision, and accuracy if average distance to plot center is used in estimating distance to center of target for five shots.

A distinction is made between continuous and discrete variables because the two types of data may require different statistical procedures. Most of the sampling methods and computational procedures discussed in this book are for use with continuous variables. The procedures for discrete variables are generally more complex. Often discrete variables can be treated as continuous, especially for larger sample sizes and a large number of classes.

Distribution Functions

Distribution functions for populations show the relative frequency with which different values of a variable occur. Given such a function, the proportion of units within certain size limits can be estimated.

Each population has its own distinct distribution function but these can often be approximated by certain general types of function such as the normal, binomial, and Poisson. The bell-shaped normal distribution, familiar to most foresters, is often used when dealing with continuous variables. The binomial distribution is used with attributes. The Poisson distribution is used with counts having no fixed upper limit, particularly if zero or very low counts tend to predominate. Several of the more important distributions are described in Appendix 2.

The form of the distribution function dictates the appropriate statistical treatment of a set of data while its exact form will seldom be known. Some indications may be obtained from the sample data or from a general familiarity with the population. The methods of dealing with normally distributed data are simpler than most of the methods that have been developed for other distributions.

Fortunately, it has been shown that, regardless of the distribution of a variable, the means of large samples tend to follow a distribution that approaches the normal. This approach to normality is often used for assessing the reliability of sample-based estimates.

Tools of the Trade

Notation

In describing various sampling methods, frequent use will be made of subscripts, brackets, and summation symbols. These devices are, like the more familiar notations of +, -, and =, a concise way of expressing ideas that would be cumbersome if put into conventional language. Using and understanding them is just a matter of practice.

Subscripts—The appearance of x_i, z_{jk}, or y_{ilmn} is annoying to individuals not accustomed to them. Yet interpreting this notation is simple. In x_i, the subscript *i* means that *x* can take on different forms or values. Inserting a particular value for *i* tells which form or value of *x* we are concerned with. *i* might imply a particular characteristic of an individual and x_1 might be its height, x_2 its weight, x_3 its age, and so forth. Or the subscript might imply a particular individual. In this case, x_1 could be the height of the first individual, x_2 that of the second, x_3 that of the third, and so forth. Which meaning is intended will usually be clear from the context.

A variable (say *x*) will often be identified in more than one way. Thus, we might want to refer to the age of the second individual or the height of the first individual. This dual classification is accomplished using two subscripts. In x_{ik}, *i* might identify the characteristic (for height, *i* = 1; for weight, *i* = 2; and for age, *i* = 3) and *k* could be used to designate the individual we are dealing with. Then, $x_{2,7}$ would tell us that we are dealing with the weight (*i* = 2) of the seventh (*k* = 7) individual. This procedure can be carried to any length needed. If the individuals in the above example were from different groups we could use another subscript (say *j*) to identify the group. The symbol x_{ijk} would indicate the i^{th} characteristic of the k^{th} individual of the j^{th} group.

Summations—To indicate that several (say 6) values of a variable (x_i) are to be added together we write $(x_1 + x_2 + x_3 + x_4 + x_5 + x_6)$ or, somewhat shorter, $(x_1 + x_2 + ... + x_6)$. The three dots (…) indicate that we continue to do the same thing for all the values from x_3 through x_5.

The same operation can be expressed more compactly by

$$\sum_{i=1}^{6} x_i .$$

In words this tells us to sum all values of x_i, letting *i* go from 1 up to 6. The symbol $\sum$ is the Greek letter sigma, indicating that a summation should be performed. The *x* tells what is to be summed and the numbers above and below $\sum$ indicates the limits over which the subscript *i* will be allowed to vary. If all of the values in a series are to be summed, the range of summation is frequently omitted from the summation sign giving $\sum_i x_i$, $\sum x_i$, or sometimes $\sum x$. All of these imply that we would sum all values of x_i. The same principle extends to variables that are identified by two or more subscripts. A separate summation sign may be used for each subscript. Thus, we might have

$$\sum_{i=1}^{3} \sum_{j=1}^{4} x_{ij} .$$

This tells us to add up all the values of x_{ij}, j having values from 1 to 4 and i from 1 to 3. Written the long way, this means $(x_{11} + x_{12} + x_{13} + x_{14} + x_{21} + x_{22} + x_{23} + x_{24} + x_{31} + x_{32} + x_{33} + x_{34})$. As for a single subscript, when all values in a series are to be summed, the range of summation may be omitted, and sometimes a single summation symbol suffices. The above summation might be symbolized by $\sum_{i,j} x_{ij}$, $\sum x_{ij}$, or maybe even $\sum x$. If a numerical value is substituted for one of the letters in the subscript, the summation is performed by letting the letter subscript vary but holding the other subscript at the specified value. As an example,

$$\sum_{j=1}^{4} x_{3j} = (x_{31} + x_{32} + x_{33} + x_{34}), \text{ and } \sum_{i=1}^{5} x_{i2} = (x_{12} + x_{22} + x_{32} + x_{42} + x_{52})$$

Analogously,

$$\sum_{i \neq j}^{3} y_i y_j$$

indicates that we want to sum both i and j from 1 to 3 but omit the values when i = j.

The sum in long hand then is

$$y_1 y_2 + y_1 y_3 + y_2 y_1 + y_2 y_3 + y_3 y_1 + y_3 y_2 .$$

Brackets—When other operations are to be performed along with the addition, bracketing may be used to indicate the order of operations. For example,

$$\sum_i x_i^2$$

tells us to square each value of x_i and then add up these squared values. But

$$\left(\sum_i x_i \right)^2$$

indicates to add all the x_i values *and then* square the sum.

The expression

$$\sum_i \sum_j x_{ij}^2$$

tells us to square each x_{ij} value and then add the squares. But

$$\sum_i \left(\sum_j x_{ij} \right)^2$$

indicates that for each value of *i* we should first add up the x_{ij} over all values of *j*. Next, this $\left(\sum_j x_{ij}\right)$ is squared and these squared sums are added up over all values of *i*. If the range of *j* is from 1 to 4 and the range of *i* is from 1 to 3, then this means

$$\sum_{i=1}^{3}\left(\sum_{j=1}^{4} x_{ij}\right)^2 = \left(x_{11}+x_{12}+x_{13}+x_{14}\right)^2 + \left(x_{21}+x_{22}+x_{23}+x_{24}\right)^2 + \left(x_{31}+x_{32}+x_{33}+x_{34}\right)^2 .$$

The expression

$$\left(\sum_i \sum_j x_{ij}\right)^2$$

tells us to add up the x_{ij} values over all combinations of *i* and *j* and then square the total. Thus,

$$\left(\sum_{i=1}^{3}\sum_{j=1}^{4} x_{ij}\right)^2 = \left(x_{11}+x_{12}+x_{13}+x_{14}+x_{21}+x_{22}+x_{23}+x_{24}+x_{31}+x_{32}+x_{33}+x_{34}\right)^2 .$$

Where operations involving two or more different variables are to be performed, the same principles apply.

$$\sum_{i=1}^{3} x_i y_i = x_1 y_1 + x_2 y_2 + x_3 y_3$$

but,

$$\left(\sum_{i=1}^{3} x_i\right)\left(\sum_{i=1}^{3} y_i\right) = \left(x_1+x_2+x_3\right)\left(y_1+y_2+y_3\right).$$

Note that $\sum_i x_i^2$ is not usually equal to $\left(\sum_i x_i\right)^2$.

Similarly, $\sum_i x_i y_i$ is not usually equal to $\left(\sum_i x_i\right)\left(\sum_i y_i\right)$.

Factorials—For convenience we use the following mathematical notation for factorials, i.e.,

$$n! = n(n-1)(n-2)...1 \text{ where } n \text{ is an integer and where } 0!=1.$$

Characterizing a distribution using measures of central tendency and dispersion

The distribution of values for a population variable is characterized by constants or parameters such as the mean and the variance. The measure of central tendency gives some idea of the typical or middle value of the distribution of a variable. The principal measures used are the mean, median, and mode. Measures of dispersion indicate how much heterogeneity there is in the distribution of the variable. They summarize the degree to which values of the variable differ from one another. The most common ones used are the variance or its square root, the standard deviation, and the range.

Measures of central tendency—Probably the most widely known and used population parameter is the mean. Given a sample where all units have the same probability of selection, the population mean is estimated by

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

with sample size n and y_i the value for the variable on sample unit *i*. For example, if we have tree heights for 5 out of 10 trees with heights 20, 20, 25, 30 and 35 m, then our estimated mean height for the 10 trees is $\bar{y} = \frac{20+20+25+30+35}{5} = 26$ m.

Other estimators of central tendency can be useful occasionally too. For example, the median is the value so that half are larger and half are smaller than this value. In this example the median would be 25. The mode is the most common value occurring in the data set, which would be 20 in this case. The median finds some utility in an estimate of central tendency for highly skewed populations, the classical one being income of people. Clearly the fact that there is a number of people in the USA, for example, that have incomes of several million dollars a year and others with less than $10,000 makes the sample mean rather a poor indicator of central tendency; the median would be more appropriate. Similarly in a stand generated by seed trees, the presence of some huge diameter seed trees may make the median a more meaningful estimate as a measure of the central tendency for such a stand. If interest is in identifying beetle-infested stands where only recently attacked trees may be saved, it may be desirable to identify stands where such trees are the most common and the mode would be the best indicator of that. Johnson (2000) gives a detailed description of the above three measures of central tendency plus several others. We focus on the mean and the corresponding total $Y = N\bar{y}$ in this book, where *N* is the total number of sample units in the population.

A measure of dispersion—Although there are several measures, we will only discuss the variance or its square root, the standard deviation, because it is used by far the most in statistics.

In any population, such as a stand of trees, the characteristic of interest will usually show variation. For example, there will be variation in tree height. Older trees will be considerably taller than younger ones and both will vary from an overall mean stand height. More observations would be needed to get a good estimate of the mean height of a stand where heights vary from 2 to 80 m than where the range is from 10-15 m. The measure of variation most commonly used by statisticians is the *variance*.

The variance of a population characteristic such as tree height is a measure of the dispersion of individual values about their mean. A large variance indicates wide dispersion and a small variance little dispersion. This variance is a population characteristic (a parameter) and is usually denoted by σ^2. Most of the time we do not know the population variance so it has to be estimated from sample data.

For most types of measurements, the estimate of the variance from a simple random sample is given by

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

where s^2 is the sample estimate of the population variance, and $\bar{y}$ is the arithmetic mean of the sample, as defined in (1) above. Sometimes, computation of the sample variance is simplified by rewriting the above equation as

$$s^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = \frac{\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}{n-1}. \tag{3}$$

Suppose we have observations on three units with the values 7, 8, and 12. For this sample our estimate of the variance is

$$s^2 = \frac{\left(7^2 + 8^2 + 12^2\right) - \frac{27^2}{3}}{2} = \frac{257 - 243}{2} = 7\ .$$

Note that the units on variance are the square of the units of the observations. If interest is in height in meters (m) then the variance will be in m^2. If interest is in tree volume in m^3 then the variance would be in m^3 squared. To avoid puzzlement we will not show the units of the variances. Also, we do not distinguish between population values Y_i and sample values y_i. It has been our experience that this distinction is unnecessary and confusing for the objectives of this book**.**

The *standard deviation* is the square root of the variance and is expressed in the same units as the mean and the variable. It is symbolized by *s*, and in the above example would be estimated as $s = \sqrt{7} = 2.6458$. Both the terms "variance" and "standard deviation" are used extensively in statistics.

Standard errors and confidence limits

Sample estimates are subject to variation just like the individual units in a population. The mean diameter of a stand as estimated from a sample of three trees will frequently differ from that estimated from other samples of three trees in the same stand. One estimate might be close to the mean but a little high, another might be much higher, and the next might be below the mean. The estimates vary because different units are observed in the different samples. And one would expect that generally, a sample of size six would generate better estimates than a sample of size three. It is desirable to have some indication of how much variation might be expected among sample estimates. An estimate of mean tree diameter that would ordinarily vary between 11 and 12 cm would inspire more confidence than one that might range from 7 to 16 cm, even though the average is the same. As discussed above, the variance and the standard deviation ($\sigma = \text{standard deviation } = \sqrt{\text{variance}}$) are measures of the variation among individuals in a population. Measures of the same form are used to indicate how a series of estimates might vary. They are called the variance and the standard error of the estimate $\sigma_{\bar{y}} = \text{standard error of the estimate of } \bar{y} = \sqrt{\text{variance of the estimate of } \bar{y}}$). The term "standard error of estimate" is usually shortened to *standard error* when the estimate referred to is obvious.

The standard error is merely a standard deviation but among estimates rather than individual units. In fact, if several estimates were obtained by repeated sampling of a population, their variance and standard error could be computed from equation (3) above. However, repeated sampling is unnecessary; the variance and the standard error can be obtained from a single set of sample units. Variability of an estimate depends on the sampling method, sample size, and variability among the individual units in the population, and these are the pieces of information needed to compute the variance and standard error. For each of the sampling methods described later on, a procedure for computing the standard error of estimate will frequently be given.

Computation of a standard error is necessary because a sample estimate may be meaningless without some indication of its reliability. If it takes 100 birds of a rare species to maintain or grow its population in a forest, we may feel good when the manager tells us that he estimates there to be 150. But how useful is that information? If we subsequently find out that the actual estimate is between 0 and 300, we have a much more realistic picture of the real situation and realize that we still do not know whether the population will survive or not and that we have to obtain better information. Figure 2 from Czaplewski (2003) illustrates the importance of a good sample size in constructing confidence intervals.

Given the standard error of estimate, it is possible to estimate limits that suggest how close we might be to the parameter being estimated. These are called confidence limits. For large samples we can take as a rough guide that, unless a 1-in-3 chance has occurred in sampling, the parameter will be within one standard error of the estimated value. Thus, for a sample mean tree diameter of 16 cm with a standard error of 1.5 cm, we can say that the true mean is somewhere within the limits 14.5 to 17.5 cm. In making such statements we will, over the long run, be right on average two times out of three. One time out of three we will be wrong because of natural sampling variation. The values given by the sample estimate plus or minus one standard error are called the 67-percent confidence limits. By spreading the limits we can be more confident that they will include the parameter. Thus, the estimate plus or minus two standard errors will give limits that will include the parameter unless a 1-in-20 chance has occurred. These are called the 95-percent confidence

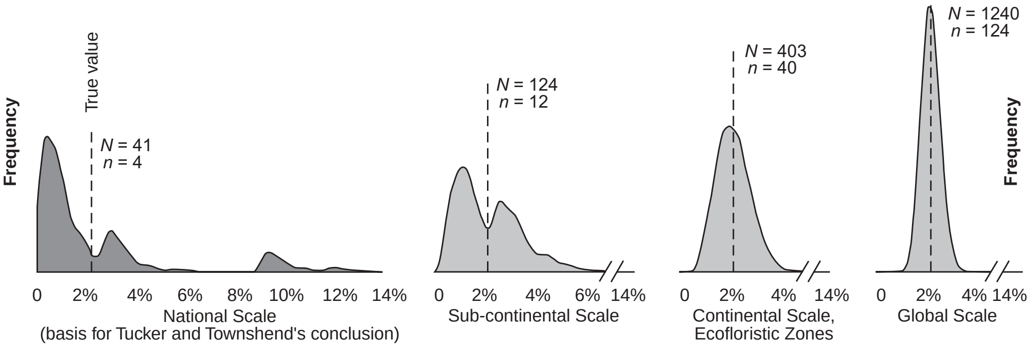


Figure 2. Estimated extent of tropical deforestation with a 10 percent sample of Landsat satellite scenes (Czaplewski 2003).

limits. The 99-percent confidence limits are defined by the mean plus or minus 2.6 standard errors. The 99-percent confidence limits will include the parameter unless a 1-in-100 chance has occurred.

It must be emphasized that this method of computing confidence limits will give valid approximations only for large samples. The definition of a large sample depends on the population itself but, in general, any sample of less than 30 observations would not qualify. Some techniques of computing confidence limits for small samples will be discussed later for a few of the sampling methods.

Expanded variances and standard errors

Very often an estimate is multiplied by a constant to generate estimates of other parameters, for example going from an estimate of the mean to an estimate of the total for a population. If a survey has been made using one-fifth ha plots and the mean volume per plot computed, this estimate would be multiplied by 5 in order to express it on a per-ha basis, or, for a tract of 800 ha, multiplied by 4,000 to estimate the total volume.

Since expanding an estimate in this way must also expand its variability, it will be necessary to compute a variance and standard error for these expanded values. This is easily done. If the variable *y* has variance s^2 and this variable is multiplied by a constant (say *k*), the product (*ky*) will have a variance of k^2s^2.

Suppose the estimated mean volume per one-fifth ha plot is $14\,m^3$ with a variance of the mean of 25 giving a standard error of $\sqrt{25} = 5m^3$. The mean volume per ha is: 5(14) = 70 m^3 and the variance of this estimate is $5^2 \times 25 = 625$ with standard error $\sqrt{625} = 25m^3$.

Note that if the standard deviation of *y* is *s* or the standard error of $\overline{y}$ is $s_{\overline{y}}$, then the standard deviation of *ky* is *ks* and the standard error of $k\overline{y}$ is $ks_{\overline{y}}$. This makes sense since constants have no variability. So, in the above case, since the standard error of the estimated mean per one-fifth ha is 5, the standard error of the estimated mean volume per ha is 5×5 = 25. A constant may also be added to a variable. Such additions do not affect variability and require no adjustment of the variance or standard errors. Thus if $z = y + k$ with *y* a variable and *k* a constant, then $s_z^2 = s_y^2$. This situation arises where for computational purposes the data have been coded by the subtraction of a constant. The variance and standard error of the coded and uncoded values are the same. Suppose we have the three observations 127, 104, and 114. For ease of computation, these could be coded by subtracting 100 from each, to make 27, 4, and 14. (This was an important advantage in the past when computers had limited capabilities and had trouble dealing with very large values especially when used in computing variances.) The variance of the coded values is:

$$s^2 = \frac{\left(27^2 + 4^2 + 14^2\right) - \frac{45^2}{3}}{2} = 133,$$

the same as the variance of the original values

$$s^2 = \frac{\left(127^2 + 104^2 + 114^2\right) - \frac{345^2}{3}}{2} = 133.$$

Coefficient of variation

The coefficient of variation, *C*, is the ratio of the standard deviation to the mean.

$$C = \frac{s}{\overline{y}}, \tag{4}$$

Thus, for a sample with a mean of $\overline{y} = 10$ and a standard deviation of $s = 4$,

$$C = \frac{4}{10} = \text{ 0.4 or 40 percent}.$$

Variance, our measure of variability among units, is often related to the mean size of the units; large items tend to have a larger variance than small items. For example, the variance in a population of tree heights would be larger than the variance of the heights of a population of shrubs. The coefficient of variation expresses variability on a relative basis. The population of tree heights might have a standard deviation of 4.4 m while the population of shrubs might have a standard deviation of 0.649 m. In absolute units, the trees are more variable than the shrubs. But, if the mean tree height is 40 m and the mean height of the shrubs is 5.9 m, the two populations may have the same relative variability, i.e., a coefficient of variation of $C = 0.11$.

Variance also depends on the measurement units used. In our example above, the standard deviation of shrub heights was 0.649 *m*. Had the heights been measured in dm, the standard deviation would have been 10 times as large (if $z = 10y$, $s_z = 10s_y$) or 6.49 *dm*. But the coefficient of variation would be the same regardless of the unit of measure. In either case, we would have

$$C = \frac{s}{\overline{y}} = \frac{0.649m}{5.9m} = \frac{6.49dm}{59dm} = \text{ 0.11 or 11 percent}.$$

In addition to putting variability on a comparable basis, the coefficient of variation simplifies the job of estimating and remembering the degree of variability of different populations. In many of the populations with which foresters deal, the coefficient of variation could be 100 percent or more. Because it is often possible to guess at the size of the population mean, we can roughly estimate the standard deviation. Such information is useful in planning a sample survey.

Covariance, correlation, and regression

Covariance and correlation are measures of how two variables vary in relationship to each other (covariability). In some sampling applications two or more variables are measured on each sample unit. In measuring forage production, for example, we might record the green weight of the grass clipped to a height of 1 cm from a circular plot 1 m in diameter. Later we might record the ovendry weight of the same sample. We would expect that there would be a positive relationship between these two variables.

Suppose the two variables are labeled *y* and *x*. If the larger values of *y* tend to be associated with the larger values of *x*, the covariance will be positive. If the larger values of *y* are associated with the smaller values of *x*, the covariance will be negative. When there is no particular association of *y* and *x* values, the covariance approaches zero. Like the variance, the covariance is a population characteristic, a parameter.

For simple random samples, the formula for the estimated covariance of *x* and *y* (s_{xy}) is

$$s_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{n-1} \tag{5}$$

Computation of the sample covariance is simplified by rewriting the formula

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{n-1}. \tag{6}$$

Suppose a sample of n = 6 units produced the following x and y values, say green weight and ovendry weight of the forage example above:

"i"	1	2	3	4	5	6	Totals
y	2	12	7	14	11	8	54
x	12	4	10	3	6	7	42

Then,

$$s_{xy} = \frac{(2\times 12)+(12\times 4)+...+(8\times 7)-\frac{54\times 42}{6}}{6-1} = \frac{306-378}{5} = -14.4.$$

The negative value indicates that the larger values of y tend to be associated with the smaller values of x. Clearly we should be dubious about this result and examine more carefully what happened since one would expect larger values of green weight with larger values of ovendry weight.

The magnitude of the covariance, like that of the variance, is often related to the size of the unit values. Units with large values of x and y tend to have larger covariance values than units with smaller x and y values. A measure of the degree of linear association between two variables that is unaffected by the size of the unit values is the simple correlation coefficient. A sample-based estimate of the correlation coefficient, R, is

$$r_{xy} = \frac{\text{covariance of x and y}}{\sqrt{\text{variance(x)}\times\text{variance(y)}}} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y}. \tag{7}$$

The correlation coefficient can vary between –1 and +1. As in covariance, a positive value indicates that the larger values of y tend to be associated with the larger values of x. A negative value indicates an association of the larger values of y with the smaller values of x. A value close to +1 or –1 indicates a strong linear association between the two variables. Correlations close to zero suggest that there is little or no linear association.

For the data given in the discussion of covariance we found $s_{xy} = -14.4$. For the same data, the sample variance of x is $s_x^2 = 12.0$, and the sample variance of y is $s_y^2 = 18.4$. Then the estimate of the correlation between y and x is

$$r_{xy} = \frac{-14.4}{\sqrt{12.0\times 18.4}} = \frac{-14.4}{14.86} = -0.969.$$

The negative value indicates that as x increases y decreases, while the nearness of r to –1 indicates that the linear association is very close. In this example we would become even more suspicious of the results and hypothesize for example that sample labels were switched somehow, since we would expect a strong positive relationship between green and ovendry weight.

An important thing to remember about the correlation coefficient is that it is a measure of the *linear* association between two variables. A value of r close to zero does not necessarily mean that there is no relationship between the two variables. It merely means that there is not a good linear (straight-line) relationship. There might actually be a strong nonlinear relationship.

Remember that the correlation coefficient computed from a set of sample data is an estimate, just as the sample mean is an estimate. Like the sample mean, the reliability of a correlation coefficient increases with the sample size (see Appendix 3, Table 5).

Regression analysis deals primarily with the relationship between variables of interest and other variables considered to be covariates. The idea is to use information on the covariates to improve estimation for the variables of interest either because information on the covariates is available or can be collected more readily/cheaply than on the variables of interest. For this reason we establish a linear relationship between the variable of interest y and the covariate x such that

$$y_i = \alpha + \beta x_i + e_i \text{, i=1,...,N} \tag{8}$$

where $e_i, i = 1, ..., N$ are the residuals for the population with the average residual over the population denoted by $E(e_i)$, where

$$E(e_i) = 0$$

and the covariate of residuals i and j is denoted by $E(e_i e_j)$ where $E(e_i e_j) = \sigma^2 v_i$ if $i = j$ or $E(e_i e_j) = 0$ otherwise; α and β are regression coefficients that are estimated from the data so that we can predict the y_i for the x_i that were not sampled as well as estimate the mean or total for the variable y; $\sigma^2 v_i$ denotes the variance of y at x_i (v_i is often represented as a function of x_i such as $v_i = x_i^k$). The value k is usually assumed known where k = 0 denotes a constant variance and k = 1 or 2 are often used when the variance of y_i is expected to increase linearly with some function of x_i). σ^2 is usually estimated from the data.

Independence

When no relationship exists between two variables they are said to be independent; the value of one variable tells us nothing about the value of the other. The common measures of independence (or lack of it) are the covariance and the correlation coefficient. As previously noted, when there is little or no association between the values of two variables, their covariance and correlation approach zero (but keep in mind that the converse is not necessarily true; a zero correlation does not prove that there is no association but only indicates that there is no strong *linear* relationship).

Completely independent variables are rare in biological populations, but many variables are weakly related and may be treated as if they were independent for practical purposes. As an example, the annual height growth of pole-sized loblolly pine dominants is relatively independent of the stand basal area within fairly broad limits (say 12 to 30 m^2 /ha). There is also considerable evidence that periodic cubic volume growth of loblolly pine is poorly associated with (i.e., almost independent of) stand basal area over a fairly wide range.

The concept of independence is also applied to sample estimates. In this case, however, the independence (or lack of it) may be due to the sampling method as well as to the relationship between the basic variables. Two situations can be recognized: two estimates have been made of the same parameter or estimates have been made of two different parameters.

In the first situation, the degree of independence depends entirely on the method of sampling. Suppose that two completely separate surveys have been made to estimate the mean volume per ha of a forest. Because different sample plots are involved, the estimates of mean volume obtained from these surveys would be regarded as statistically independent. But suppose an estimate has been made from one survey and then additional sample plots are selected and a second estimate is made using plot data from both the first and second surveys. Since some of the same observations enter both estimates, the estimates would be dependent. In general, two estimates of a single parameter are not independent if some of the same observations are used in both. The degree of association will depend on the proportion of observations common to the two estimates.

Problem: Two random samples of size n are taken without replacement from a population. By pure chance, the two samples are identical. Are the 2 estimates independent?

Answer: Yes, they are.

Problem. In the above example, how would you go about combining the two samples?

Answer. The most sensible solution would probably be to treat it as a sample of size 2*n* with replacement even though each sample in itself was taken without replacement. The advantage of this is that the variance estimate would typically be an overestimate of the actual variance.

In the second situation (estimates of two different parameters) the degree of independence may depend on both the sampling method and the degree of association between the basic variables. If mean height and mean diameter of a population of trees were estimated by randomly selecting a number of individual trees and measuring both the height and diameter of each tree, the two estimates would not be independent. The relationship between the two estimates (usually measured by their covariance or correlation) would, in this case, depend on the degree of association between the height and diameter of individual trees. On the other hand, if one set of trees were used to estimate mean height and another for estimating mean diameter, the two estimates would be statistically independent even though height and diameter are not independent when measured on the same tree.

A measure of the degree of association between two sample estimates is essential in evaluating the sampling error for estimates from several types of surveys. Procedures for computing the covariance of estimates will be given when needed.

Variances of products, ratios, and sums

Earlier, we learned that if a quantity is estimated as the product of a constant and a variable (say $Q = kz$, where k is a constant and z is a variable) the variance of Q will be $s_Q^2 = k^2 s_z^2$. Thus, to estimate the total volume of a stand, we multiply the estimated mean per unit ($\bar{y}$, a variable) by the total number of units (N, a constant) in the population. The variance of the estimated total will be $N^2 s_{\bar{y}}^2$. Its standard deviation (or standard error) would be the square root of its variance or $N s_{\bar{y}}$.

The variance of a product—In some situations the quantity in which we are interested will be estimated as the product of two variables and a constant. Thus

$$Q_1 = kyx \qquad (9)$$

where:

k = a constant and
y and x = variables having variances s_y^2 and s_x^2 and covariance s_{xy}.

For large samples, the variance of Q_1 is estimated by

$$s_{Q_1}^2 = Q_1^2 \left(\frac{s_y^2}{y^2} + \frac{s_x^2}{x^2} + \frac{2s_{xy}}{xy} \right) = k^2 \left[x^2 s_y^2 + y^2 s_x^2 + 2xys_{xy} \right]. \qquad (10)$$

As an example of such estimates, consider a forest survey project which uses a dot count on aerial photographs to estimate the proportion of an area that is in forest ($\bar{p}$), and a ground cruise to estimate the mean volume per ha ($\bar{v}$) of forested land. To estimate the forested area, the total land area (N) is multiplied by the estimated proportion forested. This in turn is multiplied by the mean volume per forested ha to give the total volume. In formula form

$$\text{Total volume} = N\,\bar{p}\,\bar{v}$$

where:

N = the total land area in ha (a known constant),
$\bar{p}$ = the estimated proportion of the area that is forested, and
$\bar{v}$ = the estimated mean volume per forested ha.

The variance of the estimated total volume would be

$$s^2 = (N\overline{pv})^2 \left(\frac{s_{\bar{p}}^2}{\bar{p}^2} + \frac{s_{\bar{v}}^2}{\bar{v}^2} + \frac{2s_{\overline{pv}}}{\overline{pv}} \right).$$

If the two estimates are made from separate surveys, they are assumed to be independent and the covariance set equal to zero. This would be the situation where $\bar{p}$ is estimated from a photo dot count and $\bar{v}$ from an independently selected set of ground locations. With the covariance set equal to zero, the variance formula would be

$$s^2 = (N\overline{pv})^2 \left(\frac{s^2_{\bar{p}}}{\bar{p}^2} + \frac{s^2_{\bar{v}}}{\bar{v}^2} \right).$$

Variance of a ratio—In other situations, the quantity we are interested in is estimated as the ratio of two estimates multiplied by a constant. Thus, we have

$$Q_2 = k\frac{y}{x}. \tag{11}$$

For large samples, the variance of Q_2 can be approximated by

$$s^2_{Q_2} = Q^2_2 \left[\frac{s^2_y}{y^2} + \frac{s^2_x}{x^2} - \frac{2s_{xy}}{xy} \right] \tag{12}$$

as is still often used, for example, by Freese (1962) or Cochran (1977). A more robust estimator is

$$v_J(\hat{Y}_{rm}) = \frac{N^2(1-f)\bar{X}^2(n-1)\sum_{j=1}^{n} D^2_{(j)}}{n} , \tag{13}$$

where $f = n/N$, $\bar{X}$ is the population mean for variable x , and for every j in the sample $D_{(j)}$ is the difference between the ratio $\frac{(n\bar{y}-y_j)}{(n\bar{x}-x_j)}$ and the average of these n ratios (Schreuder and others 1993).

Variance of a sum—Sometimes we might wish to use the sum of two or more variables as an estimate of some quantity. With two variables we might have

$$Q_3 = k_1x_1 + k_2x_2 \tag{14}$$

with k_1 and k_2 constants and x_1 and x_2 variables having variance s^2_1 and s^2_2 and covariance s_{12}.

The variance of this estimate is

$$s^2_{Q_3} = k^2_1s^2_1 + k^2_2s^2_2 + 2k_1k_2s_{12}. \tag{15}$$

If we measure the volume of sawtimber (x_1) and the volume of poletimber (x_2) on the same plots (and in the same units of measure) and find the mean volumes to be $\bar{x}_1$ and $\bar{x}_2$, with variances s^2_1 and s^2_2 and covariance s_{12}, then the mean total volume in pole-size and larger trees would be $\bar{x}_1$ + $\bar{x}_2$. The variance of this estimate is

$$s^2 = s^2_1 + s^2_2 + 2s_{12}. \tag{16}$$

The same result would, of course, be obtained by totaling the *x* and *y* values for each plot and then computing the variance of the totals. This formula is also of use where a weighted mean is to be computed. For example, we might have made sample surveys of two tracts of timber.

Example:

Tract 1
Size = 3,200 ha
Estimated mean volume per ha = 48 m^3
Variance of the mean = 11.25

Tract 2
Size = 1,200 ha
Estimated mean volume per ha is 74 m^3
Variance of the mean = 12.4

In combining these two means, to estimate the overall mean volume per ha, we might want to weight each mean by the tract size before adding and then divide the sum of the weighted means by the sum of the weights. This is the same as estimating the total volume on both tracts and dividing by the total acreage to get the mean volume per ha. Thus

$$\overline{x} = \frac{3200(48)+1200(74)}{3200+1200} = 55.09 \,.$$

Because the two tract means were obtained from independent samples, the covariance between the two estimates is zero, and the variance of the combined estimate would be

$$s_{\overline{x}}^2 = \left(\frac{3200}{4400}\right)^2 (11.25) + \left(\frac{1200}{4400}\right)^2 (12.40) = \frac{(3200)^2(11.25)+(1200)^2(12.40)}{(4400)^2} = 6.8727 \,.$$

The general rule for the variance s_Q^2 of a sum

$$Q = k_1x_1 + k_2x_2 + \ldots + k_nx_n = \sum_{i=1}^{n} k_ix_i \tag{17}$$

is

$$\begin{aligned} s_Q^2 &= k_1^2s_1^2 + k_2^2s_2^2 + \ldots + k_n^2s_n^2 + 2k_1k_2s_{12} + 2k_1k_3s_{13} + \ldots + 2k_{n-1}k_ns_{(n-1)n} \\ &= \sum_{i=1}^{n} k_i^2s_i^2 + \sum_{i \neq j}^{n} k_ik_js_{ij} \end{aligned} \tag{18}$$

where:

> $k_i, i = 1,\ldots,n$ are constants, x_i are variables with variances s_i^2 and covariance s_{ij}, for $i = 1,\ldots,n$ and $j(\neq i) = 1,\ldots,n$.

Transformation of variables

Some of the statistical estimation procedures described already and in later sections imply certain assumptions about the nature of the variable being studied. When a variable does not fit the assumptions for a particular procedure, some other method must be used or the variable must be changed to fit the assumptions or, as we say in statistics, transformed.

One of the common assumptions is that variability is independent of the mean. Some variables (e.g., those that follow a binomial such as proportion of trees that are of a particular species or Poisson distribution such as a count of number of trees) generally have a variance that is in some way related to the mean, i.e., populations with large means often having large variance. In order to use procedures that assume that there is no relationship between the variance and the mean, such variables are frequently transformed. The transformation, if properly selected, puts the original data on a scale in which its variability is independent of the mean. Some common transformations are the square root, arcsin, and logarithm.

If a method assumes that there is a linear relationship between two variables, it is often necessary to transform one or both of the variables so that they satisfy this assumption. For example the relationship between total tree volume and dbh is usually curvilinear whereas the relationship between volume and dbh squared is usually linear. A variable may also be transformed to convert its distribution to the normal on which many of the simpler statistical methods are based. Good discussions on transformations are given in Kutner and others (2003) and Carroll and Rupert (1988).

Finally, note that transformation is not synonymous with coding (say dividing all numbers by 1,000), which is done to simplify computation. Nor is it a form of mathematical magic aimed at obtaining answers that are in agreement with preconceived notions. But interpretation of results becomes more complicated with transformations. We might understand a relationship between number of birds per ha and the density of a desirable plant species, but explaining a linear relationship between log (number of birds) and log (plant density) is hard to grasp even if the latter is required for valid statistical estimation. When possible, estimates should be transferred back to the original scale of interest. This is not always straightforward as can be seen in the references cited above.

IV. Sampling Strategies

Designs With the Horvitz-Thompson Estimator

We discuss only single-phase probability sampling designs in this chapter, i.e., we assume that there is a sampling frame available from which we can select a sample directly. This could be a sampling frame of trees, field plots, recreation users, campgrounds, or sampling days for recreation use. Recall that a sampling strategy comprises both the sampling design and the estimator(s) used.

For clarity of understanding, we start with the simplest probability design: simple random sampling (SRS) to illustrate the concepts underlying probabilistic sampling. This is combined with the simple estimator of the total or the mean of the variable of interest to give us a simple sampling strategy. This allows us to point out that the simple mean and total estimators are special cases of the general unbiased unequal probability sampling estimator, called the Horvitz-Thompson estimator. We then go on to look at the general case of unequal probability sampling and note how SRS, stratified sampling, cluster sampling, sampling with probability proportional to size (pps), and to some degree systematic sampling with a random start are special cases and why they are good sampling designs to use under specific circumstances.

In the text a small population of size 10 is used with the data shown in Table 1. For computer oriented readers, Appendix 4 uses a more realistic large mapped population called Surinam with some worked examples. Readers can use the examples in the text and others directly with program R as shown in that Appendix. This data set consists of a 60 ha stem-mapped population of trees from a tropical forest in Surinam used and described by Schreuder and others (1997). This population of 6,806 trees has the relative spatial location of the trees and can be used to illustrate the efficiency of several sampling strategies.

Simple random sampling (SRS)—This is the simplest probabilistic approach. All samples of size n (samples including n sample units) have the same probability of selection. All sample units have probability of selection *n/N* and each set of two units have joint probability of selection $\frac{n(n-1)}{N(N-1)}$ in the most usual situation of sampling without replacement. This may appear to be difficult to implement because there are $\frac{N!}{n!(N-n)!}$ possible samples if sampling is without replacement (so that all *n* units are distinct). For example, for a small population of size 10 as in Table 1 with two units selected, there are 10!/(2! 8!) possible distinct samples. Selecting distinct units is more efficient than selecting with replacement, where a unit can be selected and measured more than once. This should be intuitively reasonable, since remeasuring a unit already in the sample does not provide us any new information as would the measurement of a new unit for the sample. Tied to this is the concept of the finite population correction (fpc) = $\frac{N-n}{N}=1-\frac{n}{N}$, based on the sampling fraction (*n/N*) = f. The fpc is usually part of the variance estimate and indicates that as sample size *n* goes to

Table 1. A small population used for illustration of some of the ideas discussed, where *y* = variable of interest and x_1 and x_2 are covariates.

Unit	Age	*y* = tree volume	x_1 = basal area (ba)	x_2 = remotely measured ba
1	5	1	1	2
2	5	2	2	2
3	3	3	3	2
4	6	4	4	2
5	7	5	5	2
6	9	10	10	4
7	10	10	10	4
8	12	10	10	4
9	12	10	20	4
10	15	20	20	4
		***Y* = 75**	**X_1 = 85**	**X_2 = 30**

population size *N*, the variance estimate becomes zero. This is true because basically we are measuring the entire population as the sampling fraction goes to one, or stated in another way, as the fpc goes to zero. We often ignore the fpc because many populations are quite large and sample sizes are small so that the fpc is essentially 1.

SRS is not hard to implement conceptually if there is a list of the population units available. One only has to make sure that the selection of any one unit is not influenced by the others selected or to be selected. For example, one can give each of the units a distinct number from 1 to *N* and then select n distinct random numbers between 1 and *N*. Traditionally one could use a random number table (Appendix 3, Table 1) but it is often more convenient now to use a random number generator, also given in the Appendix.

SRS sampling also has the advantage that since all units have the same probabilities of selection, applicable analysis techniques are easy to implement and estimation is straightforward and understandable, e.g., when estimating the mean μ or total *Y* of a population. The unbiased estimator of the total is:

$$\hat{Y} = \frac{N\sum_{i=1}^{n} y_i}{n} \tag{19}$$

with sample size n and y_i the value for variable of interest on sample unit *i*. The variance of the sample mean is:

$$V(\bar{y}) = \frac{N^2(N-n)}{Nn}\times\frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{(N-1)} = \frac{N^2(N-n)}{N}\frac{S^2}{n} = N^2(1-f)\ \frac{S^2}{\mathrm{n}}\cdot \tag{20}$$

An unbiased estimator of the variance of the estimated total is:

$$v(\hat{Y}) = \frac{N^2(N-n)}{Nn}\times\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{(n-1)} = \frac{N^2(N-n)}{N}\frac{s^2}{n} \tag{21}$$

where:

N = number of sample units in the population and
s^2 is the sample variance.

An estimator of the mean μ, $\bar{y}$ would be obtained by dividing $\hat{Y}$ by *N*, so $\bar{y} = \hat{Y}/N$ and its variance would be $v(\bar{y}) = v(\hat{Y})/N^2$.

Example:

Assume we have the small population shown in Table 1 and are interested in estimating either the average tree volume, $\mu = \bar{Y}$, or total volume *Y*, for this mini-forest. A possible sample of size *n* = 4 is:

Sample 1

Unit	1	2	3	4
Value	1	2	3	4

Then the estimated average tree volume for the population of trees is:

$$\bar{y} = \frac{(1+2+3+4)}{4} = 2.5 \text{ and the variance is:}$$

$$s^2 = \frac{(1-2.5)^2+(2-2.5)^2+(3-2.5)^2+(4-2.5)^2}{(4-1)} = \frac{(-1.5)^2+(-.5)^2+.5^2+1.5^2}{3} = 1.67$$

$$\text{and } v(\bar{y}) = \frac{10-4}{10\times4}\times1.67 = 0.2505\,.$$

If interest is in the total *Y*, our estimate would be $\hat{Y} = 10\times2.5 = 25$ with estimated variance $v(\hat{Y}) = 100\times0.2505 = 25.05\,.$

Note that this is not a good sample since the actual $Y = 75$. But for all SRS samples, the average value of $\hat{Y}$ would be 75. To illustrate how sampling estimates can vary dramatically with SRS, take another random sample of size $n = 4$ from this population, say (1, 2, 9, 10).

Sample 2

Unit	1	2	9	10
Value	1	2	10	20

Then the estimated average timber volume for the population of trees is:

$$\bar{y} = \frac{1+2+10+20}{4} = \frac{33}{4} = 8.25$$

$$s^2 = \frac{(1-8.25)^2 + (2-8.25)^2 + (10-8.25)^2 + (20-8.25)^2}{(4-1)} = 77.58$$

and $v(\bar{y}) = \frac{10\text{-}4}{10\times 4} 77.58 = 11.64$.

$$\hat{Y} = 10\times 8.25 = 82.5 \text{ and } v(\hat{Y}) = 100\times 11.64 = 1164 .$$

Therefore, the first, inaccurate estimate shows a small estimated variance whereas the second estimate is much more accurate but shows a large estimated variance. This is something that can happen with probability sampling, especially with SRS, which is why we have other designs that typically perform much better on average.

Problem: What are the advantages of SRS? Identify at least one key drawback.

Answer: The overriding advantage of SRS is the simplicity in analysis. The equally serious disadvantage is that it often is quite inefficient in estimation since more reliable and informative probabilistic samples can usually be collected.

Note that for the simple population of size 10 above, there are $\frac{10!}{4!6!} = 210$ without-replacement samples of size 4, but 715 with-replacement samples (ignoring the order of units selected). Clearly it would be advantageous if we can improve the chances of favoring the selection of some of those samples over others in the probabilistic sampling context if more is known about the population. For example, it makes sense to have the units selected be different to gain maximum information about the population. Hence selecting a without-replacement sample is clearly better than a with-replacement sample if we note that for samples of size four there are only 210 completely distinct samples out of 715 with-replacement samples, 360 with three distinct units, 135 with two distinct units, and 10 with only one distinct unit. Hence only 210/715 = 0.34 of the with-replacement samples contain the maximum of information for 4 units in them.

Problem: Show that for large populations with small sample size, it does not make a difference whether or not with or without replacement sampling is used.

Answer: Especially for small sample sizes, the maximum information is desired for the sample taken. So a sample consisting of all different units is better than one containing duplicates. Then the probability of n distinct units in a sample of n units out of a population of N units is P(n out of n distinct) = $N(N-1)(N-2)...(N-n+1)/N^n$. For example, for a population of 10 units with a sample size of 4 this is: 5,040/10,000 = 0.504. For a population of 20 units with $n = 4$, this becomes: 0.727. For a population of 100 with a sample of $n = 4$, this becomes: 0.941. Clearly this probability is essentially 1 for large N holding $n = 4$.

Problem: A fuels researcher comes to you rather upset. He measured the depth of the organic matter layer on both a burned and an unburned stand. For random samples of size 5 on each stand, he obtained the following results: for the burned plot, average depth was 10 cm; for the unburned plot it was 8 cm. Could you explain what might have happened?

Answer: Encourage him to compute the standard errors of estimates for both estimates! It is likely that the variability was so high that one or both of the samples were unlucky samples in the sense that they either yielded estimates too high or too low for the sampled plots. The confidence limits for the two estimates are likely overlapping considerably indicating that the sample estimates are not statistically different—which would make the researcher feel somewhat better, at least, since he would certainly expect the burned area to have smaller average depth than the unburned area.

We can often do even better than simple random sampling without replacement. Sometimes, we may have complete knowledge on a covariate associated with the variable of interest for which we know all the values in the population; or we can often get these with relative ease. This information when combined with the information on the variable of interest measured on a sub-sample of the units can be used in various ways in sample selection and estimation.

Denoting by *y* = variable of interest and *x* = covariate, numerous sample selection schemes and estimators are possible.

Unequal probability sampling—One big advantage of unequal probability sampling is that, for a fixed sample size, it is a generalization of the other single-phase probabilistic procedures. Understanding the concept of unequal probability sampling greatly facilitates understanding of the other procedures and why it is advantageous to use them in certain circumstances. Let π_i be the probability of selecting unit *i* and π_{ij} the joint probability of selecting units i and j. Then the Horvitz-Thompson estimator of the population parameter *Y* is:

$$\widehat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}. \tag{22}$$

$\widehat{Y}_{HT}$ is an unbiased estimator of *Y* with variance:

$$V\left(\widehat{Y}_{HT}\right) = \frac{1}{2} \sum_{i \neq j}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{23}$$

or

$$V(\widehat{Y}_{HT}) = \frac{1}{2} \sum_{i \neq j}^{N} w_{ij} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{24}$$

with π_i the probability of selecting unit i, π_{ij} the probability of selecting units i and j, and $w_{ij} = \pi_i \pi_j - \pi_{ij}$.

Note that (19), (20), and (21) are special cases of (22), (23), and (24), respectively. In the following we will not give the actual variance for the different sampling strategies since they are all special cases of (24).

Unbiased variance estimators based on a sample are:

$$v_1\left(\widehat{Y}_{HT}\right) = \frac{1}{2} \sum_{i \neq j}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{25}$$

and

$$v_2\left(\widehat{Y}_{HT}\right) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + \sum_{i \neq j}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j}. \tag{26}$$

If sampling is with replacement (*wr*) and the probability of sample unit *i* at a single draw is p_i, then the estimated total $\widehat{Y}_{wr}$ is:

$$\widehat{Y}_{wr} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i} \tag{27}$$

with variance

$$V\left(\hat{Y}_{wr}\right)=\frac{1}{n}\sum_{i=1}^{N}p_i\left(\frac{y_i}{p_i}-Y\right)^2 \tag{28}$$

and unbiased variance estimator

$$v\left(\hat{Y}_{wr}\right)=\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}-\hat{Y}_{wr}\right)^2 . \tag{29}$$

Note that if all the $\pi_i=\frac{n}{N}$ and all $\pi_{ij}=\frac{n(n-1)}{N(N-1)}$ then equation (22) reduces to the simple mean in (19) for SRS and, similarly, unbiased variance estimators in (23) and (24) reduce to the unbiased variance estimator for SRS in (21). Let us examine equations (22) and (24) in more detail.

If $\pi_i = ky_i$ with k a constant, then $\hat{Y}_{HT}$ is constant, actually Y, and clearly $V(\hat{Y}_{HT})$ would be 0, the ideal situation. This is only an idealized condition that won't happen in practice but we can approximate it. For example, in the small population shown in Table 1, we are interested in total volume. If we can select trees proportional to their basal area then the ratios (y_i = volume for tree i)/(x_i = basal area for tree i) are essentially constant over the 10 trees so that $\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$ is close to 0. Since we are approximating volume rather well with basal area, such a procedure should be efficient, and this is confirmed in the variance computation since the variance estimates will be close to 0. Similarly, if our interest is number of trees, giving each tree an equal weight in selection is efficient and the procedure of selecting proportional to basal area would not be. Selecting as closely as possible with probability proportional to the variable of interest is the idea behind unequal probability sampling.

Stratified sampling—In this method, the population of interest is divided into subpopulations or strata of interest. In this case, the covariable x represents strata, e.g., say x = 1 represents an old-growth stratum, x = 2 pole tree stratum, x = 3 clearcut areas, and x = 4 agricultural/wooded lands. This is a simple but powerful extension of SRS. We simply implement SRS within each stratum. The idea behind stratification is fourfold:

- To provide information on subpopulations as well as the total population.
- To divide the population into more homogeneous subpopulations or strata and improve the efficiency in estimation by a more effective distribution of the sample.
- To enable one to apply different sampling procedures in different strata; e.g., sampling in the Amazon jungle is likely to be very different from that done in the pampas or other less forested areas.
- For convenience; e.g., sampling may be done from different field stations.

In situations where a population is relatively homogeneous, SRS may be more economical than stratified sampling.

An unbiased estimator of the population mean is

$$\overline{y}_{st}=\frac{1}{N}\sum_{h=1}^{k}N_h\overline{y}_h \tag{30}$$

with estimated variance of the mean

$$v(\overline{y}_{st})=\sum_{h=1}^{k}\frac{N_h^2}{N^2}\frac{(N_h-n_h)}{N_h}s_h^2 \tag{31}$$

where:

$\overline{y}_h$ = sample mean for stratum h,
k = number of strata,

and N_h and n_h are number of sample units in the population and sample respectively in stratum h.

In Table 1, if we stratified on the basis of the remote sensing variable, x_2, we might put the first 5 units in stratum 1 and the last 5 in stratum 2. It is clear that the within-strata variability is much less than that between strata. Then suppose we decided to select a sample of size 4, 2 samples in each stratum, such as units 1, 3, 8, 10. Thus we have:

Stratum 1:

$$n_1 = 2, \quad \bar{y}_1 = \frac{(1+3)}{2} = 2, \quad s_1^2 = \frac{(1-2)^2 + (3-2)^2}{(2-1)} = 2, \quad N_1 = 5$$

Stratum 2:

$$n_2 = 2, \quad \bar{y}_2 = \frac{(10+20)}{2} = 15, \quad s_2^2 = \frac{(10-15)^2 + (20-15)^2}{(2-1)} = 50, \quad N_2 = 5$$

and

$$\bar{y}_{st} = \frac{(5\times 2) + (5\times 15)}{10} = 8.5\,.$$

Then

$$v(\bar{y}_{st}) = \frac{1}{10^2}\left(5^2\,\frac{5-2}{5}\,2 + 5^2\,\frac{5-2}{5}\,50\right) = \frac{780}{100} = 7.8$$

Then $\hat{Y}_{st} = 10\times 8.5 = 85$ with variance estimate $v(\hat{Y}_{st}) = 100\times 7.80 = 780$.

Problem: Where would you and would you not use stratified sampling?

Answer: Use if interest is in different subpopulations (strata) or if strata are more homogeneous than the overall population; also, use if different sampling schemes are desirable for different parts of the population. Do not use if simplicity is desired, for example when differences in probabilities of selection are not desired. Generally stratification is desirable.

In stratified sampling, different sampling intensities can be used in each stratum. In proportional sampling the sampling intensity is proportional to the number of units in each stratum. In optimal allocation, sampling the overall variance estimated by (31) or the overall sampling cost, C, is minimized. Clearly this requires knowledge or an estimate of the within-strata variances and a cost function, so optimal allocation is usually at best an approximation (see Schreuder and others 1993 for details on proportional and optimal allocation).

Cluster sampling—This is another extension of SRS in that now clusters of (say) trees are sampled by simple random sampling. Cluster sampling is useful when no list of sample units is available or a list is costly to obtain, which is often the case with trees; and it is usually cheaper to visit and measure clusters of trees rather than individual trees as in SRS. In cluster sampling there are actually two covariates, for example, the area of each cluster which is kept equal (say 1-ha plots) and the number of trees in each cluster which is rarely known and becomes known only for the clusters sampled.

For maximum efficiency, clusters should be heterogeneous rather than homogeneous as with strata. Cluster sampling is most useful when no list of sample units is available or is very costly to obtain and the cost of obtaining observations increases as the distance between them increases. A mechanism for randomly selecting and locating the clusters must be available.

Suppose we select n out of *N* clusters at random for sampling and each cluster is measured completely for the variable of interest. Then for clusters of different sizes a biased estimator, $\bar{y}_{cl}$, of the mean per unit is:

$$\bar{y}_{cl} = \frac{\sum_{i=1}^{n} M_i \bar{y}_{i.}}{\sum_{i=1}^{n} M_i} \tag{32}$$

where M_i is the number of units in cluster *i*, with an estimator of the variance:

$$v(\overline{y}_{cl}) = \frac{(N-n)}{Nn} \frac{\sum_{i=1}^{n} \frac{M_i^2}{\overline{M}_n^2} (\overline{y}_{i.} - \overline{y}_{cl})^2}{(n-1)} \tag{33}$$

with *N* = number of clusters in the population, *n* = number of clusters selected by SRS, $\overline{M_n} = \frac{\sum_{i=1}^{M} M_i}{n}$, the average number of units per cluster in the sample, and $y_{i.}$ = the total for all observations in cluster *i*. This estimator is asymptotically unbiased, which means that as $n \longrightarrow N$, the bias goes to 0.

Using the data in Table 1 let us take a cluster sample. This is not something that can usually be done in practice but we assume it can be done here to illustrate a point. Let us first do it in an undesirable way, i.e., have minimal variability in the clusters. If we put units 1-2 in cluster 1, 3-4 in cluster 2, ..., and 9-10 in cluster 5, we would have little within-cluster variability and considerable variability between clusters. To implement cluster sampling with *n* = 4, we set up 5 clusters of 2 units each as indicated above, and select 2 of those clusters at random. If the following clusters were selected

Cluster *i*	Sample units	Volume $\overline{y}_{i.}$
1	1, 2	1.5
5	9, 10	15

then

$$\overline{y}_{cl} = \frac{(5 \times 1.5) + (5 \times 15)}{10} = 8.25$$

and

$$v(\overline{y}_{cl}) = \frac{10-2}{10 \times 2}\left[\frac{25}{25}(1.5-8.25)^2 + \frac{25}{25}(15-8.25)^2\right] = 18.225.$$

Then $\hat{Y}_{cl} = 10 \times 8.25 = 82.5$ and $v(\hat{Y}_{cl}) = 100 \times 18.225 = 1822$.

Recall that for cluster sampling we would like considerable variability within clusters. If we put units 1 and 10 in cluster 1, 2 and 9 in cluster 2, ..., and 5 and 6 in cluster 5, we would have maximum variability within the clusters. Suppose the following two clusters are now selected:

Cluster i	sample units	volume $\overline{y}_{i.}$
1	1,10	10.5
5	2, 9	6

Then

$$\overline{y}_{cl} = \frac{(5 \times 10.5) + (5 \times 6)}{10} = 8.25$$

and

$$v(\overline{y}_{cl}) = \frac{10-2}{10.2}\left[\frac{25}{25}(10.5-8.25)^2 + \frac{25}{25}(6-8.25)^2\right] = 4.55.$$

Then $\hat{Y}_{cl} = 10 \times 8.25 = 82.5$ and $v(\hat{Y}_{cl}) = 100 \times 4.55 = 455$.

Clearly, based on the results of the two samples, the second set of clusters was much more effective than the first one in efficient estimation of the mean or total volume.

Problem: Assume you wish to estimate the average age of the 10 trees in Table 1. You are allowed to core one tree in each of three clusters to determine age and you can set up the clusters as you like. How would you go about assigning trees to the three clusters? How would you assign the trees to three strata selecting one tree from each stratum randomly?

Answer: To maximize the information collected it would be best to group the 10 trees to maximize the variability within clusters for cluster sampling and to minimize the variability within groups for stratified sampling. Although no information is given on the age of the trees, it is most reasonable to assume that age is positively correlated with either volume or basal area. This means that for cluster sampling cluster 1 might be (1,2,9,10), cluster 2: (3,4,8) and cluster 3: (5,6,7). For stratified sampling: stratum 1: (1,2,3), stratum 2: (4,5,6) and stratum 3: (7,8,9,10).

pps sampling—In sampling with probability proportional to size (pps sampling), we sample proportional to the covariate (or independent variable). This is efficient when y and x are highly positively and linearly correlated. For example, basal area, x_1, is an excellent covariate when sampling for total tree volume, y. In Table 1, tree 10 would have 20 times the probability of selection of tree 1 if trees were selected proportional to basal area. The information collected on the covariate and on the variable of interest is then combined in the unbiased Horvitz-Thompson estimator to generate an estimate of, say, total volume.

It is usually best to sample without replacement rather than with replacement. The problem with pps sampling without replacement is that when the sample size is larger than 2, the joint probabilities of selection needed for variance estimation are usually not computable. There are also questions of ease of implementation, fixed sample size, and selection probabilities exactly proportional to size. Many procedures have been developed to avoid such problems in pps sampling, e.g., Brewer and Haniff (1983) discuss 50, and more have been developed since. Some of the difficulties and some of the key methods are also discussed by Schreuder and others (1993, p. 57-62). One advantage of pps sampling is that the other procedures discussed (SRS, stratified sampling, cluster sampling) are special cases of it.

An unbiased estimator of the population mean is:

$$\bar{y}_{HT} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i} \tag{34}$$

with an unbiased variance estimator:

$$v\left(\bar{y}_{HT}\right) = \sum_{i=1}^{n}\frac{\pi_i\pi_j - \pi_{ij}}{N^2\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{35}$$

where:

n = number of units in sample and
N = number of units in population.

To illustrate pps sampling, assume using Table 1, that the sample consists of units 3, 6, 9,10 selected proportional to basal area, x_1. Then:

$$\bar{y}_{HT} = \frac{85}{4\times 10}\left(\frac{3}{3}+\frac{10}{10}+\frac{10}{20}+\frac{20}{20}\right) = \frac{85}{40}\times 3.5 = 7.437 \text{ and } \hat{Y}_{HT} = 74.37\,.$$

We have not computed the variance estimate in (35) because it requires the joint probabilities of selection for the four units selected. We can compute that in this case but it is not easy. We did compute the bootstrap variance estimate as shown in Table 2.

Table 2. Comparison of results from sampling the small population in Table 1 using five sampling methods with a sample size of 4 units.

Method		Estimated total	Estimated variance of the total
SRS	case 1	25	25
	case 2	82.5	1164
Stratified sampling		85	780
Cluster sampling	case1	82.5	1822
	case 2	82.5	455
pps sampling		74.4	585
Systematic sampling		45	245

Problem: Think of a situation in natural resources sampling where pps sampling would really be efficient!

Answer: The classical answer is the selection of trees proportional to basal area if interest is in volume. This is done using a prism currently.

Generally, we would not recommend pps sampling in actual practice. In multivariate inventories, it is unlikely that there is a covariate that is positively correlated with all or several variables. Even when interest is only in one variable, often times stratified sampling can guarantee us an efficient allocation of sample units to different sizes of units. On the other hand, with pps sampling even the less desirable samples consisting of the *n* smallest or largest units are probabilistically possible.

Connectivity of the above designs—To get a feeling for when to use the above sampling strategies from a theoretical point of view, consider the variance in (24) called *V* here for convenience.

If all units have the same probability of sampling and all sets of n sample units have equal probability of selection, then with all joint probabilities of selection being equal, i.e., $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$, the weights w_{ij} are $w_{ij} = \pi_i\pi_j - \pi_{ij} = \frac{n(N-n)}{N^2(N-1)}$ for all units *i* and *j* so that all ½ *N*(*N*-1) terms in the summation contribute to the variance in (24). As noted earlier, this is simple random sampling (SRS). For SRS using the data in Table 1 with *n* = 4, all $\pi_{ij} = 12/90 = 2/15 = 0.133$ and all $w_{ij} = 4/25 - 2/15 = 2/75 = 0.027$.

In general, assuming all the π_i are equal to *n/N* and making some of the π_{ij} equal to $\frac{n^2}{N^2}$ so that the corresponding $w_{ij} = 0$, implies that those *i* and *j* have to be selected independently. For such units *i* and *j*, the π_{ij} increased from $\frac{n(n-1)}{N(N-1)}$, so some of the other π_{ij} have to be reduced correspondingly because the sum of all joint probabilities is $\sum_{i\neq j}^{N}\pi_{ij} = n(n-1)$. To reduce the variance *V*, it would be advantageous if $w_{ij} = 0$ for large values of $\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$ or equivalently $\left|y_i - y_j\right|$ for equal probabilities of selection even if this increases w_{ij} for small values. This is the idea behind stratified sampling, where we try to put units that are dissimilar into separate strata to maximize $\left|y_i - y_j\right|$ and similar units into the same ones. For example, in Table 1 if interest is in volume, we can use remotely sensed basal area x_2 as a covariate. It then makes sense if with two strata we put units 1-5 into stratum 1 and 6-10 in stratum 2 because (say) for *n* = 4, with 2 units per stratum, then $\sum_{i\neq j}^{N}\pi_{ij} = 4\times 3 = 12$ with the joint probabilities of selection of 2 units within each stratum being

$\pi_{ij}=(2/5)*1/4=1/10=0.10$ and the joint probability of 2 units in different strata being $\pi_{ij}=(2/5)(2/5)=4/25=0.16$. Then $w_{ij}=0$ for units in different strata and $w_{ij}=3/50=0.06$ for units in the same stratum. Clearly this is an effective strategy relative to SRS because we have attached the higher joint probabilities of selection to units in the same stratum (which are quite homogeneous) and the lower probabilities to units in the two separate strata.

The ideal in cluster sampling is that negative weights w_{ij} should be attached to larger values of $\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$ for reductions in *V*. No π_{ij} can exceed π_i or π_j, so that for example if all $\pi_i = n/N$ then all $\pi_{ij} \le n/N$. $\pi_{ij} = n/N$ implies that if *i* is selected then so is *j*. Thus all sample units for which $\pi_{ij} = n/N$ are all selected together. This is the idea of a cluster. To make some of the $w_{ij} < 0$, we want the π_{ij} that are equal to *n/N* to be attached to the largest differences $\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$, which implies that the members within each cluster ideally should vary as much as possible. For example, in Table 1 for estimating volume assume there are 5 clusters of size 2 each and take a sample of *n* = 4. Then, as one good option, put units 1 and 10 in cluster 1; 3 and 9 in cluster 2; 3 and 8 in cluster 3; 4 and 7 in cluster 4; and 5 and 6 in cluster 5. Then the probabilities of selection for each unit is 2/5 = 0.40 but now the joint probability of 2 units in the same cluster is $\pi_{ij}=2/5=0.40$ and in separate clusters is $\pi_{ij}=(2/5)*1/4=1/10=0.10$ so that $w_{ij}=-1/25=-0.04$ for units in the same clusters and $w_{ij}=3/50=0.06$ for units in different clusters.

For the example of *n* = 4 above using the data in Table 1, we have for SRS that all $\pi_{ij}=0.133$ with $w_{ij}=0.027$. For stratified sampling we have $\pi_{ij}=0.40$ with $w_{ij}=0$ for units in different strata and $\pi_{ij}=0.10$ with $w_{ij}=0.06$ for units within the same stratum. For cluster sampling, $\pi_{ij}=0.40$ and $w_{ij}=-0.04$ for units in the same clusters and $\pi_{ij}=0.10$ and $w_{ij}=0.06$ for units in different clusters. As the results in Table 2 show, stratification and cluster sampling can reduce the variance of the estimates dramatically relative to SRS.

The idea behind cluster sampling is antithetical to the idea behind stratification. Cluster sampling is more risky than stratified sampling. There will be sharp gains if the clusters are chosen well but sharp losses if the negative w_{ij} are associated with small values of $\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$. In stratified sampling the w_{ij} are changed much less typically than in cluster sampling because few sample units will be selected with joint probability 1 as in the latter. This is all elegantly explained in Stuart (1964).

In probability proportional to size (pps) sampling, a version of unequal probability sampling, it is assumed that there is a covariate that is positively correlated with the variable of interest, the ultimate dream being that *y* and *x* are essentially the same so that *V* is essentially 0. We do reasonably well in that respect with x_1 in regards to estimating volume *y* in Table 1 as noted before. Pps sampling is even more risky than cluster sampling. For example, if the w_{ij} are held constant, it is clear that $\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$ can be very large if the probabilities π_i are negatively correlated with the y_i.

Systematic sampling with a random start—In this type of sampling, a random sample unit is first selected as a starting unit and then every k^{th} unit thereafter is selected. Systematic sampling assumes that the population can be arrayed in some order, which may be natural—such as days of the week in recreation sampling—or artificial, such as numbered plot locations on a map. The ordering must be carefully considered in the first case but may be haphazard in the latter. For example, when sampling the use of a recreation area, we probably would not want to sample every seventh day, say every Sunday. In the past systematic sampling has not been generally endorsed by theoretical statisticians but practitioners and applied statisticians have prevailed because it is a practical way of collecting information in the field and avoids the problem of poorly distributed samples as can happen in random sampling. In general, SRS estimation procedures are used in systematic sampling (with a random start), the assumption being that the variance estimate for SRS should generally give an overestimate of the variance actually achieved with systematic sampling.

Systematic sampling with a random start should not be used when the population is distributed in a grid pattern and the sample pattern may match it! For example, in sampling recreation use of an area it may not be desirable to select every seventh day since clearly a sample consisting of every Monday could yield quite different results from a sample of every Sunday.

Problem: What is a practical situation in forestry where systematic sampling would really be efficient?

Answer: In most field situations, it is usually practically more efficient to put in a grid of plots or select a systematic sample of trees in the forest.

An unbiased estimator of the population mean is:

$$\bar{y}_{syst} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad (36)$$

with biased variance estimator:

$$v\left(\bar{y}_{syst}\right) = \frac{N-n}{N}\frac{s^2}{n}. \qquad (37)$$

Note that these formulas are the same as for SRS.

From Table 1, assume we decide to select our starting point at random from units 1, 2, and 3 and unit 2 is selected. Then, if $n = 4$, we would take units 2, 5, 8, and 1. We do this by using mode 10 numbering, i.e., we select units 2, 5, 8, and 11 so that 11 becomes 1. Then our estimate would be:

$$\bar{y}_{syst} = \frac{1+2+5+10}{4} = 4.5$$

with $$s^2 = \frac{(1-4.5)^2 + (2-4.5)^2 + (5-4.5)^2 + (10-4.5)^2}{3} = \frac{49}{3} = 16.33$$

so $$v\left(\bar{y}_{syst}\right) = \frac{\frac{10-4}{10}\times 16.33}{4} = 2.45.$$

Then

$$\hat{Y}_{syst} = 45 \text{ and } v(\hat{Y}_{syst}) = 245.$$

Problem: Assuming the population is visited in the above order with a systematic sample of size three, starting with unit 2, what samples of size three cannot occur?

Answer: One example: units 2, 3, and 4 cannot occur together.

In Table 2 we show the results from the examples above for the various sampling methods. It is clear that both estimated totals and their variance estimates vary considerably from sample to sample. Being quite inefficient in this case, one would expect SRS to vary much more than the others, and the table certainly indicates tremendous differences in the results for the two SRS samples. One would expect the other methods to vary much less. The cluster sampling results show dramatically the differences between effective clustering, as in case 2, vs. poor clustering, as in case 1 in table 2. Pps sampling should be especially efficient here since we are sampling proportional to basal area, which is quite closely linearly related to volume in this small population.

Problem: Show how stratified sampling with "optimal" allocation is an unequal probability sampling procedure. Show that even with proportional allocation it should be considered such.

Answer: In optimal sampling the units in different strata would have different probabilities of selection. In proportional allocation two units in the same stratum would typically have different joint probabilities of selection than two units in different strata.

Problem: Assume in a population of 25 grizzly bears that the meat consumption for bear 13 is typical. Bear 1 eats only ½ the average of the 25 bears, and bear 24 eats as much as the other 24 combined, according to a local wildlife specialist. She is willing and able to give you good estimates of the amount eaten by each bear. If due to a limited budget, we can sample the actual consumption of only 1 bear and we need to make sure that enough meat is provided to minimize mauling of customers, how would you pick the sample bear?

Answer: If you decided to sample proportional to estimated consumption as given by the wildlife specialist, you failed! It is clearly best to select bear 13. This is an example of using common sense rather than applying theory. One has to make an immediate decision and selecting either bear 24 or bear 1 for example would yield useless results for making such a decision. This example is a modification of a circus elephant example given by Basu (1971) to illustrate the blind use of probabilistic sampling. In his example, the statistician recommended using pps sampling and was promptly fired by the circus director for giving such bad advice.

Problem: Show how systematic sampling with a random start can be considered a special case of:
stratified sampling
cluster sampling

Answer: It is stratified sampling where one unit is selected per stratum or it can be considered cluster sampling where all selected units form a cluster.

Variance Estimation in General

Classical variance estimation was discussed earlier. The variances are typically derivable and usually give unbiased or at least consistent estimates of the actual variance. In many cases, however, the actual sampling strategy used is quite complex and such "classical" variance estimators cannot be derived and, hence, variance estimates cannot be computed. For such situations and even in cases where the actual variances can be derived and computed, other methods are available, the two best known being jackknifing and bootstrapping. We only discuss bootstrapping since it is the easiest to implement in most situations.

Bootstrapping is a clever technique taking full advantage of the computing power that we now have worldwide. This computer-based method allows one to calculate measures of precision to statistical estimates. Confidence intervals can be constructed without having to make normal theory assumptions. The basic concept is most easily understood for simple random sampling. Suppose we have a sample of n units of y, with sample mean $\bar{y}$ and variance $v(\bar{y})$. Bootstrapping is accomplished by taking a sample of n units with replacement from the n sample units. This is done B

times. Then for each of the *B* samples we compute $\bar{y}_b, b = 1, ..., B$ with average $\tilde{\bar{y}}_B$. The variance between these bootstrap estimates is:

$$v(\bar{y}_B) = \frac{\sum_{b=1}^{B} (\bar{y}_b - \tilde{\bar{y}}_B)^2}{B-1}. \tag{38}$$

This variance estimator can also be used for $\bar{y}$. In addition the *B* sample estimates generate a distribution of estimates for easy confidence interval construction. The selection of the bootstrap samples should mimic the actual sample selection method used. Using simple random with replacement bootstrap sampling from a sample selected by unequal probability sampling is unacceptable. So is the application of bootstrapping to a purposive sample. There are various ways of bootstrapping described, for example, in Schreuder and Williams (2000). When both the bootstrap and classical variance estimates can be computed it is not yet clear which is best to use. The bootstrap method yields immediate, non-symmetric confidence intervals whereas the classical variance is easier to compute.

Regression and Ratio Estimators

Although the Horvitz-Thompson estimator is efficient in many situations, it can be quite unreliable in some. For ease of understanding we limit ourselves to one covariate; those interested in several covariates should consult Sarndal and others (1992). Consider a population where some of the covariate values, *x*, are quite small relative to the values of the variable of interest, *y*. It is clear that if some of the sample units contain *y* and *x* values where *x* is quite small, these ratios in the estimator, *y/x* could be quite large. For example if *x* = 0 for one or more units, its ratio would be undefined. Units with *x* = 0 would not be selected by pps sampling (causing bias in the estimation) but would be with SRS. Having extreme ratios can cause serious problems with the mean-of-ratio estimators (only the Horvitz-Thompson one was discussed here) and their use is generally not recommended at all with SRS.

Regression and ratio-of-means estimators, like stratification, were developed to increase the precision or efficiency of a sample by making use of supplementary information about the population being studied. The critical difference of when to use the regression or the ratio-of-means estimator is illustrated in Figure 3. Consider the linear relationships between two variables *x* and *y* shown with the line marked *A* passing through the origin and the one marked *B* intersecting the ordinate *y*.

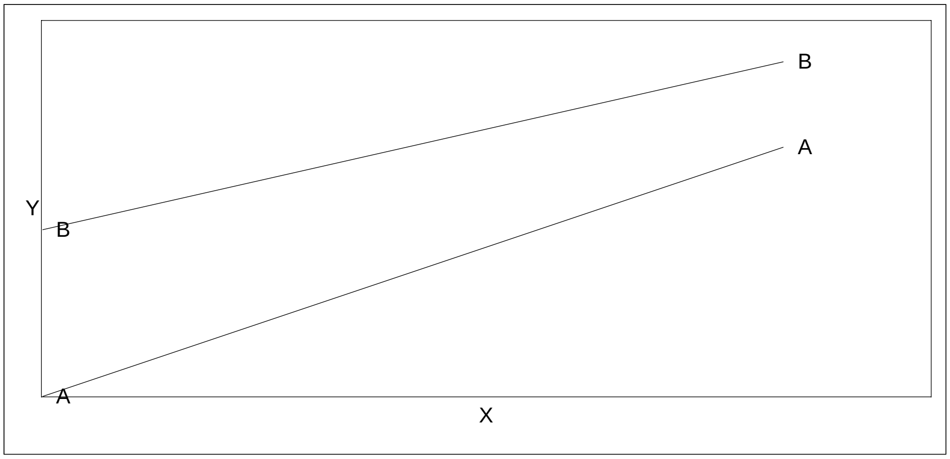


Figure 3. Postulated relationships between variables y and x.

If line *B* is the relationship expected between the variables, so that clearly the relationship does not go through the origin, one should use regression. With relationship *A* through the origin, ratio estimation is indicated. Mathematically both regression and mean-of-ratio estimators are based on the following model being reasonable for the data

$$y_i = \alpha + \beta x_i + e_i,\ i = 1,...N \text{ where } E(e_i) = 0 \text{ and } E(e_i e_j) = \sigma^2 v_i \text{ if } i{=}j \text{ and } E(e_i e_j) = 0 \text{ otherwise.} \tag{39}$$

Here $E(e_i)$ indicates the average error for the regression model over the population of *y* and *x* values, $E(e_i e_j)$ denotes the covariance of the errors given that the average error is zero and $\sigma^2 v_i$ denotes the variance of *y* at x_i (v_i is often represented as a function of x_i such as $v_i = x_i^k$ where *k* = 0 denotes a constant variance and *k* = 1 or 2 are often used when variance of y_i is expected to increase with x_i).

Then if $\alpha \neq 0$ use a regression estimator and if $\alpha = 0$ approximately use a ratio estimator. When in doubt, it is generally better to use the regression estimator. Ordinarily the question is answered based on our knowledge of the population and by special studies of the variability of *y* at various values of *x*. If we know the way in which the variance changes with changes in the level of *x*, a weighted regression procedure may be used by setting *k* to known values such as *k* = 1 or 2.

Regression estimation—Assuming a straight line relationship between y and x with constant variance (i.e., v_i = 1, *i* = 1,..., *N*) is still the most generally accepted approach at this time. The equation for the line can be estimated from

$$\bar{y}_R = \bar{y} + b\left(\bar{X} - \bar{x}\right) = a + b\bar{x} \tag{40}$$

where:

$\bar{y}_R$ = the mean value of *y* as estimated at a specified value of the variable *x*, $\bar{x}$.
$\bar{y}$ = the sample mean of *y*,
$\bar{x}$ = the sample mean of *x*,

$b = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, the linear regression coefficient of *y* on *x*, and

$a = \bar{y} - b\bar{x}$ = the intercept of *y* on *x*.

As noted in Sarndal and others (1992), the regression estimator is equal to the Horvitz-Thompson estimator plus an adjustment term. The regression estimator works well when the adjustment term is negatively correlated with the error of the Horvitz-Thompson estimator. For large errors in the Horvitz-Thompson estimator, the adjustment terms will be about equal to the errors but of the opposite sign for large samples with a strong linear relationship between the variables *y* and *x*.

Standard error for regression—In computing standard errors for simple random sampling and stratified random sampling, it was first necessary to obtain an estimate (s_y^2) of the variability of individual values of *y* about their mean. To obtain the standard error for a regression estimator, we need an estimate of the variability of the individual *y*-values about the regression of *y* on *x*. A measure of this variability is the standard deviation from regression $(s_{y,x})$ computed by

$$s_{y.x} = \sqrt{\frac{SS_y - \frac{(SP_{xy})^2}{SS_x}}{n-2}} \tag{41}$$

where $SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$, and $SP_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$.

Then the standard error of $\bar{y}_R$ is

$$s_{\bar{y}_R} = s_{y,x}\sqrt{\left(\frac{1}{n} + \frac{(X - \bar{x})^2}{SS_x}\right)\frac{N-n}{N}} \; . \tag{42}$$

So for *y* = volume and x_1 = basal area in Table 1 for a sample of *n* = 4 with observations (1,2,9,10) we have:

$Y = 0.401 + 0.73x_1$ so our estimated mean volume is
$\bar{y}_R = 0.401 + 0.73 \times 10.75 = 8.25$ and the estimated total volume is
$\hat{Y}_R = 10(8.25) = 82.5$ with standard deviation from regression
$s_{y.x} = 5.0$ and standard error from regression: $s_{\bar{y}_R} = 2.5$.

It is interesting to compare $s_{\bar{y}_R}$ with the standard error that would have been obtained by estimating the mean volume by SRS from the *y*-values only. An estimated mean volume per tree is $\bar{y} = 8.25$ with standard error of $s = 8.8$, and standard error of the estimate of $s_{\bar{y}} = 4.4$.

The family of regression estimators—The regression procedure in the above example is valid only if certain conditions are met. One of these is, of course, that we know the population mean for the supplementary variable (*x*). As will be shown in a later section (double sampling for regression), an estimate of the population mean can often be substituted. Often the *x* variable can be measured on a much larger sample than the *y*-variable so that our estimate for the *x*-variable is much better and can be used to improve the estimate for the y-variable.

The linear regression estimator that has been described is just one of a large number of related procedures that enable us to increase our sampling efficiency by making use of supplementary information about the population. Two other members of this family are the ratio-of-means estimator and the mean-of-ratios estimator. The Horvitz-Thompson estimator can be considered an example of the mean-of-ratios estimator. It is very dangerous to use with equal probability sampling such as SRS, and we will only discuss ratio-of-means estimation here.

The *ratio-of-means estimator* is appropriate when the relationship of *y* to *x* is in the form of a straight line passing through the origin and when the standard deviation of *y* at any given level of *x* is proportional to the square root of *x*. This means that in equation (39) we assume that $\alpha \doteq 0$ and that $v_i = x_i$ approximately for all *i* = 1,…,*N* units in the population. The ratio estimate $\left(\bar{y}_{rm}\right)$ of mean *y* is

$$\bar{y}_{rm} = \hat{R} \times \bar{X} \tag{43}$$

where

$\hat{R}$ = the ratio of means obtained from the sample $= \frac{\bar{y}}{\bar{x}} = \frac{\sum y}{\sum x}$ and
$\bar{X}$ = the known population mean of *x*.

The standard error of this estimate can be reasonably approximated for large samples by the jackknife variance estimator:

$$v_J\left(\hat{Y}_{rm}\right) = N^2(1-f)\bar{X}^2(n-1)\frac{\sum_{i=1}^{n} D_{(j)}^2}{n} \; , \tag{44}$$

where for every j in the sample, $D_{(j)}$ is the difference between the ratio $\frac{n\bar{y} - y_j}{n\bar{x} - x_j}$ and the average of these *n* ratios. This robust estimator often provides an overestimate of the actual variance (Schreuder and others 1993).

It is difficult to say when a sample is large enough for the standard error formula to be reliable, but Cochran (1977) has suggested that *n* must be greater than 30 and also large enough so that the ratios $\frac{s_{\bar{y}}}{\bar{y}}$ and $\frac{s_{\bar{x}}}{\bar{x}}$ are both less than 0.1.

From this sample the ratio-of-means using the same sample of four trees as for the regression estimator is:

$$\hat{R} = 33/43 = 0.77 .$$

The ratio-of-means estimator is then

$\bar{y}_{rm} = \hat{R}\ \bar{X}$ = 0.77*8.5 = 6.52 and the standard error of the estimated total is $v_J\left(\hat{Y}_{rm}\right) = 1.5$.

This computation is, of course, for illustrative purposes only. For both the regression and the ratio-of-means estimators, a standard error based on less than 30 observations is usually of questionable value.

Warning! The reader who is not sure of his knowledge of ratio and regression estimation techniques would do well to seek advice before adapting regression estimators in his sampling. Determination of the most appropriate form of estimator can be very challenging. The ratio estimators are particularly troublesome. They have a simple, friendly appearance that beguiles samplers into misapplications. The most common mistake is to use them when the relationship of *y* to *x* is not actually in the form of a straight line through the origin (i.e., the ratio of *y* to *x* varies instead of being the same at all levels of *x* or $\alpha \neq 0$). To illustrate, suppose that we wish to estimate the total acreage of farm woodlots in a county. As the total area in farms can probably be obtained from county records, it might seem logical to take a sample of farms, obtain the sample ratio of mean forested acreage per farm to mean total acreage per farm, and multiply this ratio by the total farm acreage to get the total area in farm woodlots. This is, of course, the ratio-of-means estimator, and its use assumes that the ratio of *y* to *x* is a constant (i.e., can be graphically represented by a straight line passing through the origin). It will often be found, however, that the proportion of a farm that is forested varies with the size of the farm. Farms on poor land tend to be smaller than farms on fertile land, and, because the poor land is less suitable for row crops or pasture, a higher proportion of the small-farm acreage may be left in forest. The ratio estimator may be seriously biased.

The total number of diseased seedlings in a nursery might be estimated by getting the mean proportion of infected seedlings from a number of sample plots and multiplying this proportion by the known total number of seedlings in the nursery. Here again we would be assuming that the proportion of infected seedlings is the same regardless of the number of seedlings per plot. For many diseases this assumption would not be valid, for the rate of infection may vary with the seedling density.

In general, more complex but also more robust estimators should be used. The following generalized regression and ratio-of-means estimators are generalizations of the above simple linear and ratio-of-means estimators. There are of course other estimators possible, for example regression estimators based on nonlinear relationships between *y* and *x,* but those are only applicable in very specific situations—especially since transformations may often make the relationship between variables linear so that linear regression estimation can be used on the transformed scale.

A very general efficient estimator, the generalized regression estimator (Sarndal 1980), is:

$$\hat{Y}_{gr} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} + a_{gr}\left(N - \sum_{i=1}^{n} \frac{1}{\pi_i}\right) + b_{gr}\left(X - \sum_{i=1}^{n} \frac{x_i}{\pi_i}\right) = \sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{n} \frac{e_i}{\pi_i} \qquad (45)$$

where:

$$\hat{y}_i = a_{gr} + b_{gr} x_i , e_i = y_i - \hat{y}_i ,$$

$$a_{gr} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i v_i} - b_{gr} \sum_{i=1}^{n} \frac{x_i}{\pi_i v_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i v_i}}$$

$$b_{gr} = \frac{\sum_{i=1}^{n} \frac{1}{\pi_i v_i} \sum_{i=1}^{n} \frac{x_i y_i}{\pi_i v_i} - \sum_{i=1}^{n} \frac{y_i}{v_i \pi_i} \sum_{i=1}^{n} \frac{x_i}{v_i \pi_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i v_i} \sum_{i=1}^{n} \frac{x_i^2}{v_i \pi_i} - \left(\sum_{i=1}^{n} \frac{x_i}{v_i \pi_i} \right)^2}$$

with variance:

$$V\left(\hat{Y}_{gr}\right) = \frac{1}{2} \sum_{i \neq j}^{N} \left(\pi_i \pi_j - \pi_{ij}\right) \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j} \right)^2 \qquad (46)$$

and two possible variance estimators

$$v_1\left(\hat{Y}_{gr}\right) = \frac{1}{2} \sum_{i \neq j}^{n} \frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}} \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j} \right)^2 \qquad (47)$$

and

$$v_2\left(\hat{Y}_{gr}\right) = \frac{1}{2} \sum_{i \neq j}^{n} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{e_i^{'}}{\pi_i} - \frac{e_j^{'}}{\pi_j} \right)^2 \qquad (48)$$

where:

$$e_i = y_i - \tilde{y}_s - b_{gr}(x_i - \tilde{x}_s)\,,$$

$$e_i^{'} = e_i - e_i \left[\left(\frac{(\hat{N} - N) \sum_{l=1}^{n} \frac{x_l^2}{v_l \pi_l} - (\hat{X} - X) \sum_{l=1}^{n} \frac{x_l}{\pi_l v_l}}{v_i} \right) + \left(\{ -(\hat{N} - N) \sum_{l=1}^{n} \frac{x_l^2}{v_l \pi_l} + (\hat{X} - X) \sum_{l=1}^{n} \frac{1}{\pi_l v_l} \} \right) \left(\frac{x_i}{v_i} \right) \right]$$

$$\times \left[\frac{1}{\sum_{i=1}^{n} \frac{x_i^2}{\pi_i v_i} \sum_{i=1}^{n} \frac{1}{\pi_i v_i} - \left(\sum_{i=1}^{n} \frac{x_i}{\pi_i v_i} \right)^2} \right]$$

$$,\ \hat{N} = \sum_{i=1}^{n} \frac{1}{\pi_i}, \quad \tilde{N}_s = \sum_{i=1}^{n} \frac{1}{\pi_i v_i}, \quad \hat{X} = \sum_{i=1}^{n} \frac{x_i}{\pi_i}, \quad \tilde{x}_s = \frac{\sum_{i=1}^{n} \frac{x_i}{\pi_i v_i}}{\tilde{N}_s}, \text{ and } \tilde{y}_s = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i v_i}}{\tilde{N}_s}\,.$$

Schreuder and others (1993) give some alternative variance estimators.

Problem: Show how the widely used simple linear regression estimator in (40):

$$\hat{Y}_{lr} = N(a + b\overline{X}) = \hat{Y} + b(X - \hat{X}) \text{ with } b = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ and } a = \overline{y} - b\overline{x} \text{ is a special case of } \hat{Y}_{gr}.$$

Answer: Set all $v_i = 1$ and select units by SRS, i.e., all $\pi_i = n / N$.

The generalized regression estimator in (45) takes into account both the probabilities of selection and the variance structure in the relationship between *y* and *x*. The latter is usually not known, but can often be approximated based on existing knowledge.

A generalization of the ratio-of-means estimator is:

$$\hat{Y}_{grm} = \left(\sum_{i=1}^{n} \frac{y_i}{\pi_i} \bigg/ \sum_{i=1}^{n} \frac{x_i}{\pi_i} \right) X = \left(\hat{Y}_{HT} \big/ \hat{X}_{HT} \right) X \qquad (49)$$

with approximate variance

$$V(\hat{Y}_{grm}) = V(\hat{Y}_{HT}) - 2RCov(\hat{Y}_{HT}, \hat{X}_{HT}) + R^2 V(\hat{X}_{HT}) . \qquad (50)$$

There is a good discussion on variance estimators for this ratio-of-means estimator in Schreuder and others (1993).

We recommend the use of the bootstrap variance estimator for both the generalized regression estimator in (45) and the generalized ratio estimator in (49). Both the generalized regression and the ratio-of-means estimators are biased but asymptotically unbiased in the sense that when $n \to N$, the bias goes to 0.

Problem: Show that both the generalized regression and ratio-of-means estimators are biased but asymptotically unbiased.

Answer: Proving that the estimators are biased is not easy. It requires deriving approximate formulas for the bias, something beyond the capabilities of most readers. The easiest way is to look at the formulas for the bias in books like Schreuder and others (1993). Proving that the estimators are asymptotically unbiased can be shown by letting $n \to N$ in (45) and (49). Then the sample estimators become the population parameter.

Problem. In the state of Jalisco, Mexico, all farmers of agave have to register with an industry cooperative in terms of acreage grown, when agave is planted and at what density. The cooperative wants to find out how much dies each year for each age and how much is stolen each year from the fields (agave is a very lucrative crop and each head on a harvestable plant is worth quite a bit of money). Present two alternatives to the cooperative.

Answer: We actually have a complete sampling frame of the population of interest and the solution is straightforward. We offer two possibilities:

We can stratify the population into age classes of agave and select a random sample from each stratum. Since theft should only be a problem in harvestable agave, we should take a larger sample from the harvestable age classes. In addition to the stated objectives, we might ask the cooperative if they may want the information by size of ownership too. If yes, we might impose additional stratification based on ownership and take a random sample from all such strata. Note that the disadvantage is that number of strata could easily get out of hand. If we have 9 age classes and 5 ownership size classes, we already have 45 strata. So we have a tradeoff between information by strata, each of which is presumably of interest, and possible limitations on sample size. Note that in both cases we could also use pps sampling, such as pps sampling proportional to age of the fields or size of ownership. We prefer the stratified sampling generally because the pps sampling can give undesirable sample size allocation due to random chance. We may also be able to use regression estimation rather than the Horvitz-Thompson estimator if we think some registered variable such as size of ownership might be linearly related to either mortality or incidence of theft.

Some Specific Forestry Sampling Methods

Almost all sampling methods that have proved useful in other disciplines have been used in forestry. However, only three methods unique to or of considerable interest to natural resources inventories are discussed here. For other methods see Schreuder and others (1993, 1990), and Hajek (1957). The three methods are variable radius plot sampling (VRP), fixed area plot, and Poisson sampling:

VRP sampling—This method was introduced in forestry by Bitterlich (1947) to estimate total basal area, *G*, of a forest by a simple counting technique variously known as angle count sampling, point sampling, plotless cruising, and Bitterlich sampling. The method works as follows: An assessor visits a number of locations, m, in the forest and counts the number of trees at each which, when viewed at a given height on a tree, usually breast height, subtend an angle greater than some fixed critical angle α generated by an angle gauge. This gauge could be one's thumb held at a given distance from one's eye, a simple rod with a cross piece, or for precise work, a Spiegel Relaskop or a prism. Trees are selected proportional to their cross-sectional area at the sighted point. If interest is in basal area, the trees are viewed at breast height. Since the trees are selected proportional to the variable of interest, a simple count of those selected multiplied by a known constant gives an estimate of the total basal area in the forest. In general, in analogy with equation (34), the estimator is

$$\hat{Y}_{HT} = \frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{N_k}\frac{y_{ki}}{\pi_{ki}} = \frac{\sum_{k=1}^{m}\hat{Y}_k}{m} \tag{51}$$

where $\pi_{ki} = g_{ki} / (FA)$ with g_{ki} the basal area of tree *i* at point *k*, *F* the basal area factor that determines the size of the angle α, *A* the area of the population of interest, and $\hat{Y}_k$ the estimated total basal area from point k. The variance is:

$$V(\hat{Y}_{HT}) = \left(\frac{\sum_{i=1}^{N} y_i^2}{m} + \sum_{i \neq j}^{N} \frac{y_i y_j \pi_{ij}}{\pi_i \pi_j} - Y^2 \right) \Big/ m \tag{52}$$

with an unbiased variance estimator:

$$v_1\left(\hat{Y}_{HT}\right) = \frac{\sum_{k=1}^{m}\left(\hat{Y}_k - \hat{Y}_{HT}\right)^2}{m(m-1)}. \tag{53}$$

For volume estimation the general recommendation is to select a prism (or basal area factor) resulting in a count on average of 6-10 trees at each sample point. VRP sampling has the big advantage especially to timber-oriented people that trees are selected proportional to their size and so minimizes the selection of numerous small trees.

Problem: If in VRP sampling interest is in basal area, why is the variance, *V*, not zero?

Answer: Because the sample size is random so that the variance in sample size is not zero. The variance of the basal area estimate is a combination of the variability in basal area estimates and variability in sample size. The first part is zero but the second one is not.

Problem: Several people had the idea of taking prisms of different sizes to the field and then selecting the one that gave them the desired number of trees at each point. What is wrong with this procedure? (See Schreuder and others 1981.)

Answer: It can be seriously biased. In fact, that is how it came to the authors' attention. Estimates based on the approach were so much larger than previous estimates that estimates of growth were clearly unrealistic and forest managers suspected something had gone wrong.

The basic principle used in VRP sampling is applicable in other forestry disciplines, e.g., in sampling an area for amount of recreational use. An instant count of the number of users at random times during the day gives an estimate of the amount of use for that day since users are selected proportional to their use. For example, a fisherman who is there the whole day would be counted every time a sample is taken whereas a family who spends only a few minutes would most likely be missed. Clearly, if we are interested in number of users, we need to adjust the estimated count of people by their use (i.e., their probability of selection).

Fixed area plot sampling—This procedure is usually applied using circular plots and subplots. With the general interest now in ecological as opposed to timber information, it is difficult to optimize for any specific variable in sample selection as one does with VRP sampling for volume. Because of its simplicity, fixed area plot sampling is easy to understand and implement relative to VRP sampling. In tropical areas, long rectangular plots are still often used because of ease of establishment in dense forest and rough terrain (Wood 1990).

Poisson sampling—This form of sampling, developed by Hajek (1957), was introduced into the forestry literature as 3-P sampling by Grosenbaugh (1964). Grosenbaugh proposed the method for timber sales where trees to be cut must be selected and marked and some of them can be sampled for volume at that time too. In the original application, sampling was done proportional to a covariate, which could be the ocularly estimated basal area or volume of a tree. To be efficient, the cruiser needed to be skilled. One way to implement Poisson sampling is to visit every unit i in the population and while doing that obtain the covariate value x_i for each tree (say ocular estimate of volume). Each estimate *x* is then compared to a random number generated between 0 and X / n_t where *X* is the population total for the population and n_t is the target sample size. If the random number for unit *i* is less than or equal to x_i , the unit is part of the sample to be measured; otherwise it is not. Clearly if $x_i > X / n_t$, the unit is selected with certainty. In actual implementation, *X* is not known and has to be estimated beforehand by X^* so random numbers have to be used between 0 and $L = X^* / n_t$. Here *L* is set by estimating *X* by X^* and then determining the desired sample size n_t . Wood (1988) clarifies procedures for how to select Poisson samples. Note then that the achieved sample size n_a is a random variable with variance:

$$V(n_a) = \sum_{i=1}^{N} \pi_i - \sum_{i=1}^{N} \pi_i^2 \ .$$

Hajek (1957) introduced the unbiased Horvitz-Thompson type estimator:

$$\hat{Y}_u = \sum_{i=1}^{n} y_i / \pi_i = \sum_{i=}^{n_a} \left(\frac{y_i}{n_t x_i} \right) X^* \ . \qquad (54)$$

The variance of $\hat{Y}_u$ is:

$$V(\hat{Y}_u) = \sum_{i=1}^{N} \frac{y_i^2 (1 - \pi_i)}{\pi_i} \qquad (55)$$

and an unbiased variance estimator is:

$$v(\hat{Y}_u) = \sum_{i=1}^{n_a} \frac{y_i^2 (1 - \pi_i)}{\pi_i^2} \qquad (56)$$

where n_a is the achieved sample size. $\hat{Y}_u$ is unbiased but can be a spectacularly inefficient estimator.

Grosenbaugh (1967) suggested a slightly biased but generally more efficient estimator for Poisson sampling called the adjusted estimator, $\hat{Y}_a$, where:

$$\hat{Y}_a = \frac{\hat{Y}_u n_e}{n_a} \tag{57}$$

with $n_e = X / L$, the expected sample size. An approximate variance of $\hat{Y}_a$ is:

$$V\left(\hat{Y}_a\right) \doteq \sum_{i=1}^{N} \left[p_i \left(\frac{y_i}{p_i} - Y \right)^2 / n_e \right] \left[1 + \frac{V(n_a)}{n_e^2} \right] \tag{58}$$

where $p_i = x_i / X$.

A reliable variance estimator is:

$$v\left(\hat{Y}_a\right) = \frac{X^2}{n_e} \frac{\sum_{i=1}^{n_a} \left(\frac{y_i}{x_i} - \sum_{j=1}^{n_a} \frac{y_j}{n_a x_j} \right)^2}{n_a - 1} \tag{59}$$

(Schreuder and others 1993). A special case of this where every unit has an equal probability of selection is called binomial sampling.

Problem: Show how the unbiased Poisson estimator can be very inefficient and unreliable.

Answer: Substituting $L = \frac{X^*}{n_t}$ into $\hat{Y}_u = \sum_{i=1}^{n_a} \frac{y_i}{\pi_i} = \sum_{i=}^{n_a} \frac{y_i}{n_t x_i} X^* = \sum_{i=1}^{n_a} \frac{y_i}{n_t x_i} n_t L = \sum_{i=1}^{n_a} \frac{y_i}{x_i} L$ shows that when $y_i = x_i$ for all *i* = 1,...,*N* units, our estimate can still be far from *Y* since substitution yields $\hat{Y}_u = \sum_{i=1}^{n_a} \frac{y_i}{x_i} L = \frac{n_a X^*}{n_e}$. Clearly even if on average $n_e = n_a$, our initial guess $X*$ of *Y* can often be pretty rough.

Problem: A land management agency sampled a large forest area for volume using several strata based on expected timber volume in the strata. Ten years later they wanted to resample the forest for volume and change in volume but had lost track of the probabilities of selection used earlier. They would like to treat their original sample as a simple random sample from the forest and remeasure those same plots for both volume and change in volume. Is this advisable? (See Schreuder and Alegria 1995.)

Answer: No! The referenced paper derives a formula for the bias of this procedure. It can be quite severe. An important lesson is to save the probabilities of selection of units for future use in case a random sample of these plots are to be revisited for remeasurement.

Sample Size Determination

The most frequently asked statistical question by users of sample surveys is, what sample size do I use? A first step is to specify well-defined objectives for the sampling. More money has been wasted because a person has poorly defined objectives. This often leads to unmet objectives with the sample collected. Once clear objectives are specified, the decision about sample size is much easier to make. In general, the recommendation will be to take the largest sample possible consistent with the money available. If this is not a satisfactory answer, a systematic statistical approach is called for. Typically one wants confidence intervals of a certain acceptable width to estimate a parameter *Y*, i.e., we would like a confidence interval:

$$P\left(\hat{Y} - \frac{zS_y}{\sqrt{n}} \le Y \le \hat{Y} + \frac{zS_y}{\sqrt{n}} \right) = 1 - \alpha$$

where z is the standard normal percentile, to ensure a high probability $(1-\alpha)$ and $\frac{S_y}{\sqrt{n}}$ is the standard error of estimate of the estimate $\hat{Y}$ we would like to generate. This equation implies that the parameter of interest Y is likely to be within the interval on average $(1-\alpha)\times 100\%$ of the time. The problem is that usually we do not know what S_y is and since we also do not know the sample size, the t distribution rather than the z distribution should be used. To estimate sample size, do as follows for SRS:

- Develop an equation that expresses n in terms of the desired precision of estimate. For SRS, $n \geq \frac{t_\alpha^2 s_y^2}{\lambda_t^2}$ where n is the desired sample size, t_α is the $1-\alpha/2$ quantile of the central t distribution with n-1 degrees of freedom that can readily be found in t-tables (Appendix 3, Table 2), s_y^2 is the estimated variance for variable of interest y, usually based on a preliminary sample of some sort, and $2\lambda_t = \frac{2t_\alpha s_y}{\sqrt{n}}$ is the desired width of the confidence interval specified.
- Estimate the unknown population parameters in the equations used to estimate the desired sample size. If this is not possible, a rule of thumb is to take a sample of size 50.
- Set priorities on the objectives of sampling. For example, if you have more than one characteristic of interest in the population, compromise is probably required to determine the “optimal” sample size desired to satisfy the different requirements. Is tree mortality as important as volume, etc.?
- Ensure that the value of n chosen is consistent with the resources available to take the sample. Often n is determined solely on this basis and it may well be that if one goes through the above exercise, one may recommend not sampling at all because the feasible sample size is too small. Usually this recommendation is ignored.

Problem: A research group wants to sample pollutants in the air above a fire using an airplane. The group has a budget of $2,000. You estimate that to make a reliable estimate, it takes a sample of size 50 to sample carbon dioxide and 60 to sample nitrogen. The group can only afford a sample of 1 to sample both carbon dioxide and nitrogen and is also interested in another 5 chemicals. What would you recommend?

Answer: The sensible answer is to recommend not sampling at this time until more money is available. The more likely outcome is that the group will actually do the sampling. A situation very similar to this was actually encountered by the senior author. One could argue that with the tremendous variability one can expect in this situation that a sample of size one could be worse than not sampling at all since the sample of size one could often generate a very misleading estimate of the actual parameters to be estimated.

Example: We are interested in estimating needle length on a tree with a confidence interval of no more than 10 mm at the 95 percent confidence level. Based on a small sample from another tree nearby we estimated mean leaf length to be $\bar{y} = 19.8$ and s = 4.1 mm. To achieve our objective then we need $n = \frac{t_{.05}^2 s^2}{(10/2)^2} = 2.69$. Hence we would probably take a sample of 3 needles from the tree to ensure that the sample obtained is sufficient and hope that the preliminary sample on which we based our sample size determination was valid for our tree of interest.

Problem. An organization tells you that for a population of 100,000 ha it found that a sample of size 40 ha was enough to give a reliable estimate for a given variable. It wants you to sample 10,000 ha for this variable and wants you to take a sample of size 4 since it is 1/10 of the original population and hence in its opinion should give an equally precise estimate for the smaller population. Do you agree?

Answer: No, you should not! The result is liable to be much less reliable for the smaller population. See for example Czaplewski (2003) for an actual example of a similar situation. See also Table 2.

What approach of locating plots and what types of plots should be used? The aim in sampling is to obtain a representative sample of the population of interest. Frequently in large-scale surveys, sampling is based on a grid sample with a random start. Strictly speaking, this is not a random sample since some locations will have joint probabilities of selection of 0. But it is justified as being SRS since the estimator is unbiased and the variance estimator for systematic sampling with a random start will usually be an overestimate of the variance assuming SRS.

It is likely that at some point in the future, multi-resource inventories will require different plot sizes and shapes for different variables of interest but sharing the same plot centers. But this is not true currently where often sampling for resources other than timber is superimposed on traditional timber surveys. For example, in the USA, FIA uses four circular 0.017 ha (1/24 acre) subplots sampling a 1-ha (2.5-acre) plot for most ecological tree variables and use transects for down woody materials and understory variables.

Plot and transect sampling techniques—Unbiased estimates of forest population parameters can be obtained from any combination of shape and size of sample units if done properly but the optimum combination varies with forest condition. The shapes of fixed area plots in forestry are commonly rectangular, square, circular, and narrow-width rectangular with the circular plot being by far the most common. Clusters of plots are often more efficient than single plots and are used commonly in forestry. If there is a clear gradient, rectangular plots laid out across it are efficient (remember that cluster sampling is more efficient if clusters are heterogeneous) but orientation should be decided in the office prior to sampling.

Rectangular and square plots are usually laid out by starting with a corner point located by survey (compass and tape) using an aerial photo or map. The second corner is then located and at both corners, right angles are established to locate corners three and four.

Circular plots are defined by the plot center and radius. Establishing a circular plot is usually simpler than other plot types because distances from the plot center have to be checked only for those trees within a peripheral strip of width approximately 1.5 to 3.0 m. The length of the strip and hence the number of boundary trees increases with increase in the radius of the plot. Sometimes exact measurements are needed to determine whether a tree is in or out of the plot.

Narrow rectangular plots are most convenient if information on topography and forest composition is also required as part of the survey and if dense undergrowth or difficult terrain necessitates spending a large amount of time on plot establishment. The width of the strips, determined beforehand in the office, usually ranges from 5 to 40 m depending on sampling intensity, topography, forest composition, density of undergrowth, and variability and value of the forest.

For a given sample intensity, a strip survey may be faster than a survey based on plots because the ratio of working time on the units to traveling time between them is greater for strips. Strips and plots may be combined in what are called “line plots.” With these, topographical and forest-type data are gathered on the strips and quantitative information on the forest is obtained from plots located at intervals along their length.

In forestry three procedures have been popular for sampling timber attributes such as volume, growth, mortality, etc.:

- Variable radius plot (VRP) sampling usually consisting of a cluster of four or five VRP subplots sampling a certain area such as an acre or ha. This is a version of unequal probability sampling where trees are selected proportional to basal area. It is efficient for sampling for volume and basal area, since tree basal area is of course highly correlated with volume. VRP sampling was invented by W. Bitterlich, an Austrian forester, in the 1930s although he did not publish his work until the 1940s presumably because of the intervening war. This method is still used in quite a few European countries. In the USA, the Chief of the FS mandated that the procedure not be used anymore by FIA. But this is clearly still a highly desirable procedure for a timber sale and for some other uses.

- Fixed area plot sampling. Generally a large plot is subsampled by a cluster of small circular plots. Trees are selected with equal probabilities. This is now used by FIA and NFS of the USFS and by several European countries. Rectangular plots could also be used but are not popular at this time although they might be highly desirable in tropical regions or in conjunction with remote sensing.
- Line intercept or line intersect sampling. This is used often for down woody material on the ground and understory vegetation. For the former, the inclusion probability is $l_i \sin w_i / L$ where l_i is the length of the log, w_i the acute angle between the log and the survey transect, and L the spacing between the lines.

FIA and the current vegetation system (CVS) plots used by Region 6 (Oregon and Washington) of the USFS (Max and others 1996) are compact, sampling a circular 1-ha plot. Although they can be established in the field faster than long rectangular plots, they are less efficient for estimation because of spatial correlations and the similarity of adjacent compact subplots. Measuring them duplicates much of the work already done and yields relatively little new information. Long subplots spread out over the observation area reduce the effect of spatial correlation relative to circular or square subplots sampling the same size area.

To increase the precision of the estimates for large areas, one seeks to make the plot estimates as similar as possible. To do this, one includes as much of the variability as possible within the plot, thus increasing efficiency. However, long rectangular or large square plots have a large perimeter that increases the number of decisions required on whether trees on the boundary are “in” or “out.” Long plots are advantageous for remote sensing, especially low-level aerial photography and videography. Numerous tree and stand variables, e.g., stocking (trees/ha) and mortality can be measured with a high degree of reliability using remotely sensed imagery. However, sampling subplots on the ground is desirable at this time to verify the remote sensed measurements and adjust them by regression estimation if necessary.

Characteristics of plot types used in the USA are summarized in Table 3.

The following is an overview of the advantages of different subplot sizes and shapes (Schreuder and Geissler 1999):

- Long rectangular plots are advantageous for low altitude photography measurements and plant biodiversity estimates.

Table 3. Characteristics of plot types.

Plot/subplot	FIA	CVS	40 x 250 m	25 x 400 m	20 x 500 m
Plot					
Area(ha)	1.000	1.000	1.000	1.000	1.000
Radius/dimensions (m)	56.42	56.42 m	40 x 250 m	25 x 400 m	20 x 500 m
Perimeter (m)	354.5	354.5	580	850	1040
Large subplot					
Area (ha)	0.1012	0.0763	0.1000	0.1000	0.1000
Radius/dimensions (m)	17.95	15.58	25 x 40	25 x 40	20 x 50
Perimeter (m)	112.8	97.89	130	130	140
Medium subplot					
Area (ha)	0.0168	0.020	0.020	0.020	0.020
Radius/dimensions (m)	7.32	8.02	10 x 20	10 x 20	10 x 20
Perimeter (m)	46.0	50.4	60	60	60
Small subplot					
Area	0.001	0.004	0.020	0.020	0.001
Radius/dimensions (m)	1.78	3.57	10 x 20	10 x 20	2 x 5
Perimeter (m)	11.2	22.4	60	60	14

- Rectangular plots are easier to fly and interpret, and a 1-ha plot is a convenient size to fly and photointerpret.
- Long narrow plot or transects are desirable to assess plant biodiversity (species richness and identification of species) because one wishes to cover as many habitat conditions and as large an area as possible to find rarer species. Boundary issues are relatively less important because one only has to check to see if the occasional species not found in the subplots is in or out of the plot.
- Circular subplots are advantageous for VRP sampling and for measuring other variables where boundary issues are important, as in regeneration subplots.
- Transects are advantageous for traversing a large area to measure scattered or rare objects such as woody debris on the ground.
- A series of small area samples such as soil cores are best for certain destructive and expensive measurements such as cores for assessing soil quality and soil series measurements.
- Plot designs for animals are more general than for plants. The series of articles ending with Schwarz and Seber (1999) hint at the possibility that with increasing technological improvement, animal populations may be sampled some day with the same ease as plant populations. Radio tagging, recording devices, and traps can simplify animal sampling and are often needed. Birds and large mammals cover large areas because of their mobility so sampling for them requires large plots. Slow moving animals such as worms, snails, ants, and many insects can be sampled on microplots similar to those used for plants described above but are often hard to observe and traps may be required to find them. Birds are particularly difficult to sample because they migrate so their populations are also reflected by conditions elsewhere. Counts of birds are also influenced by season, the time of the day, and weather.

Problem: You are charged with developing a sampling strategy for the states of Chiapas in Mexico and Colorado in the USA to estimate timber volumes in those states. What kind of ground plot(s) would you recommend in the field?

Answer: Chiapas has considerable tropical forest with difficult travel conditions. It is likely that long narrow plots, say 5 m x 100 m, might be best there. In Colorado travel in the forests would be easier and VRP plots may be the best way so that trees are selected proportional to basal area.

Edge Effects When Sampling at Stand Boundaries

Randomly selected plots may fall close to a stand boundary, and part of such plots may cross over into a different stand. These boundary plots have been dealt with in many ways, even to the point of moving the plots away from the boundary or entirely eliminating them. Some practices can seriously bias stand estimates, particularly in long skinny stands or fragmented landscapes where there is a lot of edge. Trees along the edge may grow very differently in diameter and form, for example where the bordering area is open, so ignoring boundary conditions is clearly wrong. Irregular shaped boundaries can introduce additional problems. For a complete technical treatment of the issues, see Schreuder and others (1993), sec 7.11.3, and Iles (2003), chapter 14.

In a practical application, probably the most commonly used and accepted method to deal with boundary plots is the mirage plot (Avery and Burkhart 1983, p. 221). To use the mirage technique, place the plot without bias where it would fall, and if part of the plot falls outside the stand boundary, install a mirage plot. From the original plot center, tally all of the trees on the plot that are also within the stand boundary. Measure the distance from the plot center to the boundary and install the mirage plot the same distance on the other side of the boundary. Tally all of the trees on the mirage plot that are within the stand boundary. In effect, the area of the plot that exists outside the stand boundary is mirrored back inside the stand boundary, resulting in counting some trees twice from points that are orthogonal projections of (l_{1s}, l_{2s}) across the stand boundaries that truncate the area of inclusion a_i. Formally the mirage method works as follows:

A sampling location (l_{1s}, l_{2s}) is randomly located within an area A. If r_i is the distance between this location and tree i and R_i is the limiting distance for being included in the sample, then $R_i = \frac{dbh_i}{2\sqrt{F}}$ or $R_i = R$. Depending on whether VRP or fixed area circular plots are used, unit u_i is included in the sample if $r_i \leq R_i$. The inclusion area a_i is a circular area concentric with u_i but truncated by the area boundary if it is within R_i of the tree. The weight attached to y_i is an integer multiple of $A/a_i(0)$ where the multiplier depends on whether u_i can also be tallied.

The mirage method has problems with irregular boundaries and with inaccessibility, for example cliffs, swamps, water, or trespass. For such areas, a method called walkthrough (Ducey and others 2004) has been introduced to address these shortcomings. For trees between the plot center and the boundary, measure the distance from the plot center to the tree center. Following along the same line, measure that same distance from the tree center to the boundary. If you are outside the boundary, the tree is counted twice; otherwise, only once. The advantage is that you never need to cross the boundary or worry about irregular shaped boundaries. A disadvantage may be that defining the boundary for each tree can be even more subjective often than for plots.

Design Issues

The following design issues are critical:

- Collect data on explanatory/stress variables such as rainfall deficiency, low soil moisture, exposure to pollution, etc. This type of data cannot usually be collected on plots but are essential in building reliable models.
- Simplicity in design. This provides flexibility over time and ease in analysis.
- Consistency of design over time. This simplifies change estimation and identifying possible cause-effect hypotheses.
- Flexibility to address new environmental or management issues while maintaining design consistency.
- Flexibility to incorporate new measurement technologies while maintaining design consistency.
- Ability to treat each sample unit as a population. This is important for example in classifying each sample to estimate acreage in forest types. This means, for example, no missing data for a sample unit because of the design used. Of course this is frequently not feasible.
- Use interpenetrating sampling or similar methods so sampling intensity can be readily increased in time and space if needed. This is a nice feature of annualized inventories if handled properly.
- Provide flexibility to accommodate replacement of plots to deal with damage caused by the measurement process (for example, trampling or destructive sampling) or denial of access to plots by private landowners—for example, sampling with partial replacement.
- Ability to handle missing data such as plots being inaccessible or landowners denying access (as noted by C. Kleinn, inaccessibility may also be caused by land mines or wildlife such as elephants and lions). Inaccessibility is best handled by setting aside a separate stratum for such plots and clearly stating the estimated size of that stratum and how estimates if any are generated for it.
- Implement a strong quality assurance program so that true changes in sample plots over time will not be confounded with changes in measurement error or subtle details in measurement protocol.
- Consider use of several plot designs at the same sample locations. Although this complicates data collection, it may well be required when a large suite of parameters is of interest. For example, for number of trees and total basal area of trees, very different plot designs are efficient: fixed area and VRP plots, respectively.

Instrumentation

Measurement techniques are covered in great detail in Schreuder and others (1993), Chapter 7. This section will serve as a supplemental update to that chapter. Although the instruments used today by the forest practitioner are very different than in the past, the underlying principles remain the same. In general, measurements are taken for the easily measured lengths and angles, and basic trigonometric relationships are used to calculate the harder to measure elements. Technological advances in electronics allow these measurements to be made easily, quickly, and accurately. In addition, rugged handheld computers allow not only capturing these measurements, but also auditing and processing them.

New diameter measurement tools—The tool of choice for most remains the d-tape or caliper. Two new tools however provide for convenience: the electronic caliper from Haglof and a new electronic diameter measurement device, functionally equivalent to the Relaskop, from Laser Technology. The caliper looks like the traditional beam caliper, but it also has a digital readout of the diameter as well as a data recorder; after a day's field work, the data is downloaded to a computer for processing. A promising new instrument, although not yet available commercially, is the electronic diameter measurement device. A lighted bar is superimposed on the tree, and the width of the bar is manipulated with the controls to coincide with the diameter of the tree. A distance is entered either manually or captured from a connected laser distance device. The distance to the tree, together with the width of the bar, allows the diameter to be calculated internally. With this instrument's 2X magnification and vertical angle encoder, it can also be used for upper stem diameters.

New height measurement tools—The key to determining height is an accurate measurement of horizontal distance to the tree. Laser distance measurement devices have proven themselves to be very effective over the past few years. Laser Technology, Newcon Optik, LaserAce, Handlaser, Opti-Logic, and others offer laser distance measurement. As with any new technology that is continually changing, search the World Wide Web for the latest information. Some have built in vertical angle encoders, and along with the internal logic they can display the height. An optional, add-on flux-gate compass is available for some models.

Another recent addition to the practitioner's toolbox is the Haglof Vertex Hypsometer, an ultrasonic distance measuring device. This system has two parts, a transponder and the hypsometer; the transponder can be placed at the plot center or hung on a tree, and then the hypsometer is used to determine the distance to the transponder, and optionally a vertical angle. Distance and height are displayed on the screen. The problem of boundary trees, that is, those that occur at or near the boundary of a plot, always arises when establishing sample units in the field. Measurement error associated with such trees can be a source of considerable error in deriving plot estimates in forest inventory. Ultrasonic distance measuring devices should make it easier to implement the miraging or walkthrough methods described earlier for sampling boundary areas.

New data recording—Source point data collection on a handheld portable data recorder (PDR) has many advantages over handwritten forms, particularly in light of the ease of data communication between the handheld and other electronic measurement devices. Direct capture of instrument output by the PDR avoids the common input errors often encountered. Even with mechanical measurement processes, keying the data into the PDR avoids the possibility of transcription errors. In addition, the PDR can be programmed to look for missing or illogical data entry values.

As the Microsoft Windows CE platform matures, many hardware and software solutions for forestry are available as replacements for the DOS and other proprietary operating systems. There are many choices of software for cruising, scaling, and sampling. Commercial software is available through most hardware vendors, and is also available through public entities.

The ready availability of inexpensive personal data assistants (PDA) has made automated field data collection much more affordable. With the addition of a hardshell case, the PDA has become a very serviceable field unit. For production field use, however, the truly rugged units with integrated keypads are preferable.

Sampling for Coarse Woody Debris (CWD)

In CWD inventories, one may be interested in both standing and fallen woody material. Since assessment of standing live and dead trees is usually done as part of a traditional timber inventory, only the sampling of fallen woody debris is discussed here. The discussion draws heavily on the review of Stahl and others (2001). We assume interest is in total volume and number of pieces. As noted by Stahl and others, there is no obvious best way of sampling CWD. But in line with the emphasis on simplicity in this book, strip or line sampling are favored. Strip sampling is the same as the other fixed area sampling techniques discussed elsewhere and hence does not need further elaboration here except that one needs to clearly decide when a log is in or out of the sample. Usually it is best to call the log in if the center of the butt end is in the strip for both volume and number of logs estimation. One could count a log in for volume if part of the log is in but the butt center is not, but this can lead to such complications as possibly having volume with a zero estimate of number of logs. The advantage of this technique is that it is simple to implement since such plots are easily laid out generally and material on the ground is readily accessible for measurements. There are also no problems with logs not lying horizontally or how crooked the stems and branches are (the latter have to be considered for estimating number of CWD units).

In line intercept, also called line intersect sampling, all units intersected by an inventory line are sampled. Usually the lines are laid out in segments with a specific spacing and orientation. Assuming the lines are laid out in a fixed direction, the inclusion probability of selection for a unit requires measurement of the projection of the length of the unit perpendicular to the orientation of the survey line. Then the estimators for variable *y*, either total volume or number of units, is:

$$\hat{Y} = L\sum_{i=1}^{m}\frac{y_i}{l_i \sin w_i} \qquad (61)$$

where *L* is the spacing between survey lines laid out systematically across the entire population, m is the number of lines, l_i is the length of the unit, and w_i the acute angle between the unit and the survey line. If *m* lines of sizes s_i are used, then the following ratio estimator should generally be more efficient:

$$\hat{Y} = A\frac{\left(\sum_{i=1}^{m}\frac{y_i}{l_i \sin w_i}\right)}{\sum_{i=1}^{m} s_i} \qquad (62)$$

with *A* the area being sampled. A complication of this sampling design can be sample logs parallel to the direction of sampling. Such logs have a probability of selection of close to zero and as indicated earlier with the Horvitz-Thompson estimator this can create seriously inflated estimates if such logs are counted in even if they are a valid part of the sample. If they are counted out when they should have been counted in, this clearly causes a bias in estimation. See Williams and Gove (2003) for more details about the potential bias. This method has the considerable advantage in that establishing and walking a line in the field is easy but suffers from the problems of having to measure angles, having to compensate for logs not lying horizontally or for crooked stems and branches, and deciding whether logs parallel to the line of sampling are in or out. A comprehensive discussion of the theory and history behind line intersect sampling is given in Chapter 13 of DeVries (1986).

Problem: Consider strip sampling where a log is counted in for volume but not for number of logs. If part of the log is in the strip but the butt center is not, is it possible to

a. Have volume estimates but a zero count of number of logs?
b. Have a positive estimate of number of logs but with zero volume?

Answers: a. Yes b. No

Wildlife Sampling

Much of the theory of sampling finite plant populations is not applicable to sampling many wildlife populations (Schreuder and others 1993, p. 326). Many animal species are mobile and hide, making detection or measurement difficult and so sampling may affect their location. There is usually no sampling frame and probabilities of selection have to be estimated usually after the sample is drawn. The existence of a specific selection probability for an individual in the population is often mainly conceptual. As a result, sampling animal populations is usually more expensive than sampling plant populations and more statistical assumptions have to be made to make estimation possible, so errors are more likely (Burnham 1980).

The primary parameters of interest in wildlife sampling are usually population size and rates of birth, immigration, emigration, and mortality. Populations are classified often as either closed or open. A closed population is assumed to have a constant size with the same members except for known removals during a study. In an open population, births, immigrations, emigrations, and deaths can occur.

Traditionally, only a single visit is made to a primary sample unit (psu). However, it is difficult to obtain repeatable animal observations within one visit, because counts are influenced by weather, time of day, and other factors. Leaving recording equipment in the field for a few weeks would enable samples to be taken at all times, day and night, and under varying weather conditions, making the observations much more repeatable. An important advantage of automatic recorders is that nocturnal and shy animals can be observed.

As noted, sampling strategies for animals are considerably more complex than for vegetation. Such devices as radio tags, classification of DNA samples from hairs and pellets encountered on sample locations, and high-detail remote sensing should make animal sampling easier in the future. Detailed procedures for sampling animal populations are given in Schwartz and Seber (1999) and Thompson and others (1998).

V. Sampling Methods for Discrete Variables

Simple Random Sampling (SRS) for Classification Data

Assume that for a population of a given rare tree species it is important to determine the proportion of male and female trees, and the sex of a tree can only be obtained easily in the fall. From a random sample of 50 trees, the number of females is 39. Then the estimate, $\overline{p}$, of the proportion that is female is:

$$\overline{p} = \text{Number having the specified attribute/Number observed} \tag{63}$$
$$= \frac{39}{50} = 0.78.$$

Standard error of estimate—The estimated standard error of $\overline{p}$ is

$$s_{\overline{p}} = \sqrt{\frac{\overline{p}\left(1-\overline{p}\right)}{n-1}\left(1-\frac{n}{N}\right)} \tag{64}$$

where: n = number of units observed.

In this example N is extremely large relative to n, and the finite-population correction (1-n/N) can be ignored, so that

$$s_{\bar{p}} = \sqrt{\frac{(0.78)(1-0.78)}{(50-1)}} = 0.05918 \,.$$

Confidence limits—For certain sample sizes, confidence limits can be obtained from Appendix 3, Table 3. In this example we found that in a sample of n = 50 trees, 39 were female. The estimated proportion of females was 0.78 and, as shown in Table 3, the 95-percent confidence limits would be 0.64 and 0.88. For samples of 100 and larger the table does not show the confidence limits for proportions higher than 0.50. These can easily be obtained, however, by working with the proportion of units *not* having the specified attribute. Thus suppose that, in a sample of n=1,000, the 95-percent confidence interval for an observed fraction of 0.22 is 0.19 to 0.25. If the true population proportion of males is within the limits of 0.19 and 0.25, the population proportion of females must be within the limits of 0.75 and 0.81.

Confidence intervals for large sample—For large samples, the 95-percent confidence interval can be computed as

$$\overline{p} \pm \left[2s_{\overline{p}} + \frac{1}{2n}\right]. \tag{65}$$

Assume that a sample of n = 250 units has been selected and that 70 of these units are found to have some specified attribute. Then,

$$\overline{p} = \frac{70}{250} = 0.280 \,.$$

And, ignoring the finite-population correction,

$$s_{\bar{p}} = \sqrt{\frac{(0.28)(0.72)}{249}} = 0.02845.$$

Then, the 95-percent confidence interval is:

$$0.280 \pm \left[2(0.02845) + \frac{1}{2(250)}\right] = 0.280 \pm 0.059 = 0.221 \text{ to } 0.339.$$

Thus, unless a 1-in-20 chance has occurred, the true proportion is between the limits 0.22 and 0.34. For a 99-percent confidence interval we multiply $s_{\bar{p}}$ by 2.6 instead of 2. (For samples of n = 250 or 1,000, the confidence interval could be obtained from Appendix 3, Table 3. For this example the table gives 0.22 to 0.34 as the limits.)

The above equation gives the normal approximation to the confidence limits. This approximation can be used for large samples. What qualifies as a large sample depends on the proportion of items having the specified attribute. As a rough guide, the normal approximation will be good if the common (base 10) logarithm of the sample size (n) is equal to or greater than

$$1.5 + 3\,(|P - 0.5|)$$

where: P = our best estimate of the true proportion of the population having the specified attribute and $|P - 0.5|$ = the absolute value (i.e., algebraic sign ignored) of the departure of P from 0.5.

Thus, if our estimate of P is 0.2 then $|P - 0.5|$ is equal to 0.3. To use the normal approximation, the log of our sample size should be greater than

$$1.5 + 3(0.3) = 2.4$$

so that n must be 251 (2.4 = log 251).

Sample size—Appendix 3, Table 3 may also be used as a guide to the number of units that should be observed in a SRS to estimate a proportion with a specified precision. Suppose that we are sampling a population in which about 40 percent of the units have a certain characteristic and we wish to estimate this proportion to within ± 0.15 (at the 95-percent level). The table shows that for a sample of size 30 with $\bar{p}$ = 0.40, the confidence limits would be 0.23 and 0.60. Since the upper limit is not within 0.15 of $\bar{p}$ = 0.40, a sample of size 30 would not give the necessary precision. A sample of n = 50 gives limits of 0.27 and 0.55. As each of these is within 0.15 of $\bar{p}$ = 0.40, we conclude that a sample of size 50 would be adequate.

If the table suggests that a sample of over 100 will be needed, the size can be estimated by

$$n = \frac{1}{\dfrac{E^2}{(4)(P)(1-P)} + \dfrac{1}{N}} \quad \text{for 95-percent confidence, and}$$

$$n = \frac{1}{\dfrac{E^2}{(6.76)(P)(1-P)} + \dfrac{1}{N}} \quad \text{for 99-percent confidence}$$

where:

E = the precision with which P is to be estimated and
N = total number of units in the population.

The table indicates that to estimate a P of about 0.4 to within $E = \pm$ 0.05 (at the 95-percent confidence level) would require somewhere between 250 and 1,000 observations. Using the first of the above formulae (and assuming N = 5,000) we find,

$$n = \frac{1}{\dfrac{(0.05)^2}{(4)(0.4)(0.6)} + \dfrac{1}{5{,}000}} = 357 \quad .$$

If we have no idea of the value of P, we will have to make a guess at it in order to estimate the sample size. The safest course is to guess a P as close to 0.5 as it might reasonably occur.

The following problem shows how dangerous it can be to sample for attributes without realizing the implications exactly.

Problem. Industry and an environmental group are arguing about how much old growth there is in a certain large forest. They agree upon the following definition of old growth: A hectare of forest is considered old growth if it contains at least one tree with a diameter breast height of 100 cm. A consultancy group is selected to make an inventory of the forest and decides to select 100 1-ha plots randomly from the forest. Because it is expensive to measure all trees on the sample plots they propose randomly selecting 4 subplots of 0.1 ha each and then classify each hectare as to whether it is old growth or not. Both industry and the environmental group want an unbiased estimate of old growth for the forest. Would they get it with this approach?

Answer: No. With this approach one can only err in one way. A hectare can be classified as not being old growth when in fact it is but it can never be classified as being old growth when it is not. Serious bias can result in such an estimate of old growth. See Williams and others (2001) for an extensive treatment of the issue involved. To obtain an unbiased estimate, all 100 1-ha plots would have to be censused.

How to select a tree or a seed at random—If we try to estimate the proportion of trees in a stand having a certain disease, we could do it by binomial sampling but this requires visiting every tree in the population and at that time determining whether it is a sample tree or not. This is SRS but is time consuming and results in a random sample size. Selecting trees completely at random then is difficult to do in a practical manner, which explains why systematic sampling with a random start is popular in such situations as a practical alternative.

In some populations, the individuals themselves are randomly located or can easily be made so. A batch of seed is such a population. By thoroughly mixing the seed prior to sampling, it is possible to select a number of individuals from one position in the batch and assume that this is equivalent to a completely random sample. Those who have sampled seed warn against mixing in such a manner that the light empty seeds tend to work towards the top of the pile. As a precaution, most samplers select samples from several places in the pile with a scoop, combine them, and treat that sample as a SRS.

Cluster Sampling for Attributes

In attribute sampling the cost of selecting and locating a unit is often very high relative to the cost of determining whether or not the unit has a certain attribute. In such situations, cluster sampling is usually preferred over SRS. In cluster sampling, a group becomes the unit of observation, and the unit value is the proportion in the group having the specified attribute.

In estimating the survival percentage of trees in a plantation, it is possible to choose individual trees for observation by randomly selecting pairs of numbers and letting the first number stand for a row and the second number designate the tree within that row. But it is inefficient to ignore all of the trees that one walks by to get to the one selected. Instead, survival counts are made in a number of randomly selected rows and averaged to estimate the survival percent if the same number of trees occur in each row. This is a form of cluster sampling, the clusters being rows of planted trees.

The germination percent of a batch of seed can also be estimated by cluster sampling. Here the advantage of clusters comes not in the selection of units for observation but from avoiding some hazards of germination tests. Such tests are commonly made in small covered dishes. If all the seeds are in a single dish, any mishaps (e.g., excess watering or fungus attack) may affect the entire test. To avoid this hazard, it is common to place a fixed number of seeds (one or two hundred) in each of several dishes. The individual dish then becomes the unit of observation and the unit value is the germination percent for the dish.

When clusters are fairly large and all of the same size, the procedures for computing estimates of means and standard errors are much the same as those described for measurement data. To illustrate, assume that 8 samples of 100 seeds each have been selected from a thoroughly mixed batch.

The 100-seed samples are placed in eight separate germination dishes. After 30 days, the following germination percentages are recorded:

Dish number	1	2	3	4	5	6	7	8	Sum of percentages
Germination (percent)	84	88	86	76	81	80	85	84	664

If p_i is the germination percent in the i^{th} dish, the mean germination percent would be estimated by

$$\overline{p} = \frac{\sum_{i=1}^{n} p_i}{n} = \frac{664}{8} = 83.0$$

The variance of $\overline{p}$ would be computed by

$$s_p^{\ 2} = \frac{\sum_{i=1}^{n} p_i^{\ 2} - \frac{\left(\sum_{i=1}^{n} p_i\right)^2}{n}}{n-1} = \frac{\left(84^2 + 88^2 + ... + 84^2\right) - \frac{(664)^2}{8}}{7} = 14.5714$$

and the standard error of $\overline{p}$ can be obtained as

$$s_{\overline{p}} = \sqrt{\frac{s_p^{\ 2}}{n}\left(1 - \frac{n}{N}\right)}$$

$$s_{\overline{p}} = \sqrt{\frac{s_p^{\ 2}}{n}} = \sqrt{\frac{14.5714}{8}} = 1.35$$, if the finite-population correction is ignored.

Here n stands for the number of clusters sampled and N is the number of possible clusters in the population. As in simple random sampling of measurement data, a confidence interval for the estimated percentage can be computed by Students t 95-percent confidence interval: $\bar{p} \pm ts_{\overline{p}}$

with t = the value of Student's t at the 0.05 level with n – 1 degrees of freedom. Thus, in this example, t has 7 degrees of freedom and $t_{.05}$ is 2.365. The 95-percent confidence interval is:

$$83.0 \pm (2.365)\ (1.35) = 83.0 \pm 3.19 = 79.8 \text{ to } 86.2.$$

Transformation of percentages–If clusters are small (less than 100 units per cluster) or if some of the observed percentages are greater than 80 or less than 20, it may be desirable to transform the percentages before computing means and confidence intervals. This is done to approximate the normal distribution better so that the confidence intervals should be more reliable. The common transformation is arcsin $\sqrt{percent}$. Appendix 3, Table 4 gives the transformed values for the observed percentages. For the data in the previous example, the transformed values are

Dish No.	**Percent**	**Arcsin** $\sqrt{percent}$
1	84	66.4
2	88	69.7
3	86	68.0
4	76	60.7
5	81	64.2
6	80	63.4
7	85	67.2
8	84	66.4
Total		**526.0**

The mean of the transformed values is $\frac{526.0}{8} = 65.75$.

The estimated variance of these values is:

$$s^2 = \frac{\left(66.4^2 + ... + 66.4^2\right) - \frac{(526)^2}{8}}{7} = 8.1486$$

and the standard error of the mean transformed value is

$$s_{\bar{y}} = \sqrt{\frac{8.1486}{8}} = \sqrt{1.0186} = 1.009$$

ignoring the finite population correction.

So the 95-percent confidence limits would be (using $t_{.05}$ for 7 df's = 2.365)

$$CI = 65.75 \pm (2.365)(1.009) = 65.75 \pm 2.39 = 63.36 \text{ to } 68.14.$$

Referring to the table again, we see that the mean of 65.75 from the acrsin transformation corresponds to a percentage of 83.1. The confidence limits correspond to percentages of 79.9 and 86.1. In this case the transformation made little difference in the mean or the confidence limits, but in general it is safer to use the transformed values even though some extra work is involved.

Other cluster-sampling designs—If we regard the observed or transformed percentages as equivalent to measurements, it is easy to see that any of the designs described for continuous variables can also be used for cluster sampling of attributes. In place of individuals, the clusters become the units of which the population is composed.

Stratified random sampling might be applied when we wish to estimate the mean germination percent of a seed lot made up of seed from several sources. The sources become the strata, each of which is sampled by two or more randomly selected clusters of 100 or 200 seeds. Similarly we might stratify a plantation into sections (strata), ones with high expected mortality and ones with lower expected mortality in order to assess survival percentage of trees by section. Two or more rows would be randomly selected in each section. In both cases not only might this be more efficient in estimating overall germination or survival percentages but we also can generate estimates for the strata, which might be of interest in their own right.

With seed stored in a number of canisters of 100 kg each, we might use two-stage sampling, the canisters being primary sample units and clusters of 100 seeds being the secondary sample units. If the canisters differed in volume (or the different sections in the plantation were of different importance), they (or the sections) could be sampled at different intensities, a form of unequal probability sampling.

Cluster Sampling for Attributes With Unequal-Sized Clusters

Frequently when sampling for attributes, it is convenient to let a plot be the sample unit. On each plot we count the total number of individuals and the number having the specified attributes. Even though the plots are of equal area, the total number of individuals may vary from plot to plot; thus, the clusters will be of unequal size. In estimating the proportion of individuals having the attribute, we definitely do not want to average the proportions for all plots because that would give the same weight to plots with few individuals as those with many.

In such situations, we might use the ratio-of-means estimator. Suppose that a pesticide has been sprayed on an area of small scrub oaks and we wish to determine the percentage of trees killed. To make this estimate, the total number of trees (x_i) and the number of dead trees (y_i) is determined on 20 plots, each 0.04-ha in size.

Plot	**No. trees (x_i)**	**No. dead trees (y_i)**
1	15	11
2	42	32
3	128	98
4	86	42
5	97	62
6	8	6
7	28	22
8	65	51
9	71	48
10	110	66
11	63	58
12	48	32
13	26	16
14	160	126
15	103	80
16	80	58
17	32	25
18	56	44
19	49	24
20	84	59
Total	1135	960
Mean	**67.55**	**48.0**

The ratio-of-means estimate of the proportion of trees killed is

$$\overline{p} = \frac{\overline{y}}{\overline{x}} = \frac{48.0}{67.55} = 0.7106\,.$$

The estimated standard error of $\overline{p}$ is

$$s_{\overline{p}} = \sqrt{\frac{1}{\overline{x}^2}\left(\frac{s_y^2 + \overline{p}^2 s_x^2 - 2\overline{p}s_{yx}}{n}\right)\left(1 - \frac{n}{N}\right)}$$

where:

$s_y^{\;2}$ = variance of individual *y* values,
$s_x^{\;2}$ = variance of individual *x* values,
s_{yx} = covariance of *y* and *x*, and
n = number of plots observed.

In this example

$$s_y^{\;2} = \frac{\left(11^2 + 32^2 + \ldots + 59^2\right) - \dfrac{960^2}{20}}{19} = 892.6316$$

$$s_x^{\;2} = \frac{\left(15^2 + 42^2 + \ldots + 84^2\right) - \dfrac{1351^2}{20}}{19} = 1{,}542.4711$$

$$s_{yx} = \frac{(11)(15) + (32)(42) + \ldots + (59)(84) - \dfrac{(960)(1351)}{20}}{19} = 1{,}132.6316.$$

With these values (but ignoring the fpc),

$$s_{\bar{p}} = \sqrt{\frac{1}{(67.55)^2}\left[\frac{892.6316+(0.7106)^2(1542.4711)-2(0.7106)(1132.6316)}{20}\right]} = 0.026 .$$

As in any use of the ratio-of-means estimator, the results may be biased if the proportion of units in a cluster having a specified attribute is related to the size of the cluster. For large samples, the bias will often be trivial.

Sampling of Count Variables

Statistical complications often arise in handling data such as number of weevils in a cone, number of seedlings on a 0.0004-ha plot, and similar count variables having no fixed upper limit. Small counts and those with numerous zeroes are especially troublesome. They tend to follow distributions (Poisson, Negative Binomial, etc.) with which it is difficult to work. If count variables cannot be avoided, the sampler's best course may be to define the sample units so that most of the counts are large and to take samples of 30 units or more. It may then be possible to apply the procedures given for continuous variables.

In order to estimate the number of larvae of a certain insect in the litter of a forest tract, 30 cm^2 samples were taken at 600 randomly selected points (Freese 1962). The litter was carefully examined and the number of larvae recorded for each sample. The counts varied from 0 to 6 larvae per plot. The number of plots on which the various counts were observed gave the following results:

Count	0	1	2	3	4	5	6	Total
Number of plots	256	244	92	21	4	1	2	600

The counts are close to following a Poisson distribution (see Appendix 2). To permit the application of normal distribution methods, the units were redefined. The new units consist of 15 of the original units selected at random from the 600. There are a total of 40 of the new units, and unit values are the total larvae count for the 15 selected observations. The values for the 40 redefined units are

14	13	16	13	13	14	15	12
16	18	11	7	9	10	11	10
12	14	13	14	14	13	9	17
15	8	12	5	13	15	13	10
12	12	20	10	9	14	15	13
						Total = 504	

By the procedures for simple random sampling of a continuous variable, the estimated mean $\bar{y}$ per unit is

$$\bar{y} = \frac{504}{40} = 12.6 .$$

The variance s_y^2 is

$$s_y^2 = \frac{(14^2+16^2+\ldots+13^2)\frac{(504)^2}{40}}{39} = 8.8615.$$

Ignoring the finite population correction, the standard error of the mean is

$$s_{\bar{y}} = \sqrt{\frac{8.8615}{40}} = 0.47.$$

The new units have a total area of 1.35 m^2 ; hence to estimate the mean number of larvae per ha the mean per unit must be multiplied by 10000/1.35 = 666.67

Thus, the mean per ha is (666.67) (12.6) = 8400.04 and the standard error of the mean per ha is (666.67) (0.47) = 313.33.

As an approximation we can say that unless a 1-in-20 chance has occurred in sampling, the mean count per ha is within the limits 8400.04 $\pm$ 2(313.33) or 7773.34 to 9026.66.

VI. Remote Sensing and Other Ancillary Information

Remote Sensing and Photography

Remote sensing can be defined as the science and art of obtaining information about objects, areas, and phenomena under investigation through analysis of data acquired by some device not in contact with these objects, areas, or phenomena (Lillesand and Kiefer 1987). Remote sensing has a number of significant advantages not attainable by ground sampling from an inventory and monitoring point of view. It provides a synoptic view of the study area, can be collected quickly over a large area, provides information about land cover in visible and nonvisible portions of the electromagnetic spectrum, is increasingly acquired and processed digitally, and provides a permanent record of the situation at the time.

Remote sensing sensors are either passive or active. Passive ones receive signals from the target itself, and active ones transmit a known signal. Passive remote sensing technologies useful in natural resource applications today include photographic and electro-optical imaging systems such as satellite borne sensors and airborne scanners. Their sensing capabilities extend from ultraviolet to well into the microwave. Active sensors including RADAR and LIDAR are just beginning to prove useful for selected applications.

Brief descriptions of the three types of sensors follow:

1. Photographic systems include camera, film, and a platform (usually an aircraft) to carry them. These systems now are often integrated with geographic positioning systems (GPS) and other electronics to help identify and record the location and position of the camera over the target to be photographed. Image resolution is primarily a function of camera lens resolution, film resolution, degradation due to image motion (forward motion, pitch and roll, and vibration), and film processing. If all goes well, a negative or reversal film positive will resolve 50 line pairs per millimeter or more. Sixty or even 70 line pairs per millimeter resolved on the film are not uncommon. Paper prints still are limited to about 25-30 line pairs per millimeter. The size of ground features resolved is a function of the above, plus the camera lens focal length and the flying height above the terrain. This results in photographs of a given scale, which, together with the image resolution resident in the system, determines what will be possible to see and interpret from the final photographs. Improvements in cameras and camera mounts, and integration of GPS with the aircraft and camera system, make pinpoint plot photography at very large scale (1:3000 to 1:1000) operationally feasible. Use of computer equipment and geographic information systems (GIS) make flight planning relatively easy and reliable. Recorded information on the photo center files makes it possible to plot out a map of a photo flight shortly after the mission is completed. Because of these acquisition and display technology improvements, very low altitude photography may have real possibilities as a principal source of information for hard-to-sample areas such as wilderness or it can at least decrease the amount of ground sampling needed. An old but still important reference on the use of photography for inventory is Aldrich (1979).

2. Electro-optical imaging sensors collect data as arrays of pixels. A pixel is defined as the smallest unit or cell of a raster image. It is usually assumed to be square in shape and consists of a digital number that represents the brightness value recorded for that pixel within a single spectral band. The ground resolution of the pixel is usually understood to be the distance that one side of the pixel represents on the ground. The key issue is to extract useful information from the spectral band data using image analysis (Holmgren and Thuresson 1998).

Such sensors include multispectral scanners, the main one of which currently is the thematic mapper (TM). The spatial resolution for TM is 30 m for six of the seven bands carried by Landsats 4 and 5. Landsat 7 has eight bands, one of which is a black and white band with 8 m resolution. Band 6, the thermal IR band, has a spatial resolution of 120 m. Advanced Very High Resolution Radiometer (AVHRR) is used regularly by the US National Oceanic and Atmospheric Administration. This was designed for daily high spatial-resolution images of regional cloud patterns for weather forecasting. The bands were designed to discriminate between clouds, water, ice, snow, and land. One band was subsequently modified on the operational sensor prior to NOAA-6 to allow also for observations suitable for vegetation studies. AVHRR is very coarse in its coverage. This is advantageous for getting a smaller number of observations for a very large area but is quite limited in its spatial resolution (from 1.1 km to a maximum of 3.5 km at nadir) for the same reason. The French system SPOT (System probatoire d'observation de la terre) is a commercial alternative to TM. It has higher resolution (10 m) but is much more expensive to acquire. Newer satellite systems with much higher resolution are now available, generally from commercial sources.

Czaplewski (1999) reviews remote sensing sources available and soon to be available for inventory purposes. He distinguishes the following categories:

- Low-resolution satellite data include AVHRR, MODIS, Orb View-2, ERS-2, and SPOT 4. Such data are inexpensive and have a 1,000-2,900 km (600-1,800 mile) swath width. Because of this wide swath, spatial resolution is poor with a pixel representing 64-128 ha (158-316 acres) in size. Such data have been useful for very large-scale maps of forested landscapes for global change models, and to detect hot spots of serious deforestation in heavily forested landscapes. But they are too coarse to reliably measure and monitor most forest conditions.
- Medium-resolution satellite data include Landsat 5 and 7, Radarsat, SPOT 2 and 4, IRS C and D, P 2 and 5, Spin 2, EOS AM-1m, and CBERS 1 and 2 with pixel sizes of 10-30 m (33 to 98 feet) wide. They are more expensive with a 50-160 km (30 to 100 mile) swath width. Such systems can separate forest from non-forest, and can identify some forest types and density classes. Landsat can generally identify clearcuts but not most partial cuts. Advanced regeneration after land clearing, urban centers, and size, shape, and connectivity of forest patches can also be measured. High quality data without clouds are generally available every 1-2 years except in humid tropical areas and many boreal forests.
- High-resolution satellites include Ikonos-2, OrbView 3 and 4, EROS B1 and B2, SPOT 5, and Quickbird 1 and 2 with 3.2- 9.6 km (2-6 mile) swath width and pixel size of from 1-3 m (3-10 feet) wide. These sensors have capabilities, limitations, and costs similar to high altitude 56.25 cm square (9-inch square) 1:40000 small scale aerial photography—as available from the USA Geological Service (USGS) national aerial photography program (NAPP), which covers an area about 8 km (5 miles) wide. Such satellite and photo data can be used to reliably distinguish some forest types, several stages of stand development, clearcuts and many partial cut areas, regeneration after land clearing, and concentrated tree mortality. Forest stands, land use, distance to adjacent roads, water bodies, forest fragmentation, and various types of urbanization can be photo interpreted.
- Large scale aerial photography with scales from 1:2500 to 1:12000 is routinely acquired by aerial survey companies for small sites. Each photo covers an area 0.16-3.2 km (0.1 to 2 miles) wide. Interpreters can reliably identify many forest cover conditions such as 10 broad forest types, 5 stages of stand development, 3 stand density classes, clearcut and partial cut areas, regeneration success rates, natural or artificial stand origin, 3-5 severity levels of tree mortality, most indicators of urbanization and fine-scale forest fragmentation, and stand size, shape, and edge measurements.

Aldrich (1979) notes that forest diseases are less easily detected and evaluated than insect damage with aerial photography because it takes a long time for visible symptoms of disease to show up. The symptoms are often not uniform over the forest and are more subtle than insect damage. Dwarf mistletoe, Dutch elm disease, oak wilt, basal canker of white pine, ash dieback, *Fomes annosus*, sulfur dioxide damage, and ozone damage are detectable with some degree of success. Large-scale color and color infrared (CIR) film (1:1584) are needed to ascertain the degree of damage, while 1:8000-1:16000 scales of CIR photography can be used to define and delineate the boundaries of the disease. 70 mm color and CIR photography can be used as part of a sampling strategy within susceptible forest types for damage assessment. Of course disturbances to the vegetation caused by windstorm, flood, fire, or human activities are relatively easy to detect on aerial photography as is change in these characteristics if the photography is repeated.

Aerial color video system imaging may have real utility especially for annualized inventories to find out more about changes observed in certain areas of special interest. Videocameras can be easily mounted on a variety of aircraft for either vertical or oblique sensing. Images can be digitized easily for computer-aided interpretation and can be used immediately since development is not needed. Video equipment is portable, versatile, easy to use, can tolerate different light conditions, and is cheaper to operate than photographic systems. Also, the operator can view the imagery on a monitor in the plane at the time of acquisition, can adjust the exposure settings interactively, and can record comments in flight. Also, the high rate of picture acquisition (30 frames/sec) provides extra data. Its disadvantages are its low spatial resolution relative to film, the difficulty of obtaining hard copy images from the data, practical limits on field-of-view because of the small tape format, difficulty in calibrating the cameras because of the automatic exposure control, and vignetting problems with the many near-IR video sensors since the camera optics are not designed for this wave length band. The value of common color video systems for natural resources and agricultural applications is limited because of the difficulty of extracting discrete spectral data from a composite video signal and the lack of spectral bands outside the visible resolution.

3. Microwave sensors are generating considerable interest at this time but the applications in forestry are still limited (see for example Lefsky and others 2002). The main data sets of non-photographic images currently comprise those collected by Landsat 1-5 and 7, SPOT, and AVHRR. These are available on computer-compatible media and in electronically reconstituted photographs. Both computer-aided media and conventional photointerpretation methods are used to interpret such data. Key references regarding inventory and monitoring using remote sensing are: Holmgren and Thuresson (1998), USDA Forest Service (1998), and Lefsky and others (2002).

Accuracy of Remotely Sensed Information

Management of lands by agencies, such as those of two federal agencies in the USA, the National Forest System of the USDA Forest Service, and the Bureau of Land Management of the U.S. Department of the Interior (USDI), requires reliable maps of variables such as percent forest cover, stand structure, and vegetation types. Such maps also require frequent updating and generating them is expensive. It is natural that remote sensing sources such as TM are used for this purpose since it is facilitated by the frequent, large-scale, digital acquisition. Considerable work has gone into making such maps. However, although TM contains useful information, the amount is limited. For example, it is unlikely to be useful for stand structure, a difficult to measure variable even on the ground. Similarly, vegetation types are difficult to define and interpretation may vary from one user to another. Ideally, TM information should be combined with geo-referenced field inventory and other mapped data to provide the necessary information for management decisions.

Remote sensing researchers desire a single coefficient to represent the accuracy of a thematic map and of each category displayed (Rosenfield and Fitzpatrick-Lins 1986). Usually the results of an accuracy assessment of a map are displayed in a matrix called a contingency table (called error matrix in remote sensing) where the columns indicate the classes defined by the standard of comparison and rows indicate the mapped ones. The elements in the contingency table are the counts in the row/column classes with the number in the last row the total count in that row class and the numbers in the last column the total count in that column class. An obvious first estimator of overall accuracy is the ratio of the sum of all correct counts over the total number of counts in the contingency table. But ideally we also want estimators of the errors of commission (the proportions of diagonal values to column sums = user’s accuracy) and of omission (the proportions of diagonal values to row sums = producer’s accuracy). A widely accepted coefficient of agreement is the Kappa statistic (K) estimated by:

$$\hat{K} = \frac{p_0 - p_c}{1 - p_c} \qquad (66)$$

where:

$$p_0 = \sum_{i,j}^{k} p_{ij} w_{ij} = \text{weighted proportion of units that agree,}$$

$$p_c = \sum_{i,j}^{k} w_{ij} p_{i.} p_{.j} = \text{weighted proportion of units with expected chance agreement, and}$$

$$p_{i.} = \sum_{j=1}^{k} p_{ij}, \; p_{.j} = \sum_{i=1}^{k} p_{ij}$$

where w_{ij} is the assigned weight of importance of agreement for (*i,j*) with $w_{ij} = 1$ for all *i,j* for the simple unweighted Kappa statistic and $0 \le w_{ij} \le 1$ for the weighted Kappa. Unequal weights can be assigned if the accuracy of some classes is more important than for others, with the disadvantage inherent in this that such weights would be subjective. Here $\hat{K} = 0$ indicates that obtained agreement equals chance agreement, $\hat{K} > 0$ indicates greater than chance agreement, $\hat{K} < 0$ less than chance agreement, and $\hat{K} = 1$ is perfect agreement.

Then user's accuracy = 1–(number correctly classified to be correct in the diagonal/number in that row) and producer's accuracy = 1–(number classified to be correct in the diagonal/number in that column).

To assess accuracy, we need a probabilistically selected sample of size n on which both truth and the map values to be assessed are available (Schreuder and others 2003). For specificity we only discuss the situations discussed in that paper here. We assume plots are used. Truth should be defined exactly for each variable and measured accordingly. It should not be defined as the best readily available information as is done frequently in remote sensing. For percent tree cover (y_1), very low altitude photography may best be used; for species composition (y_2) such photography should be combined with ground sampling; for stand structure (y_3) more emphasis is likely required on ground sampling with the photography providing some utility; for other variables such as understory vegetation especially in dense stands, reliance may have to be placed completely on ground sampling. Call the values of the variable of interest on plot i, $y_i, (i = 1, 2, 3)$ and the corresponding value of the variable on the map to be assessed for accuracy $x_i, (i = 1, 2, 3)$. For a simple random sample of *n* plots:

- We have n plots with y_i from truth coverage.
- For many variables, it is likely that some plots will contain more than one class.
- Truth may be obtained from photography alone or a combination of photo and ground information, depending on the variable of interest. However, if only photo information is used for percent cover, the truth is obtained error-free for the whole plot whereas if ground sampling is involved, the plot information may have sampling error.

Location error for the mapped information is assumed negligible since we do not know what it is. If present it will likely lead to an underestimate of the actual accuracy. Unless we have detailed information about errors in plot locations, we cannot correct for them. For a certain number of the *n* plots, all the information falls within one or more categories for the variable of interest for both the truth and the mapped information. The following treats the case of both *x* and *y* labeling only the same two "truth" classes occurring on a truth plot. The extension to more than two classes is straightforward.

For a given plot, assume that the part *x* labeled x_{ij} is part of or covers the part y_{ij} called that by the truth plot. If the truth plot and mapped plot could be overlaid completely, this assumption is not needed. But the truth plot may only provide estimated areas of the plot area in the classes of interest for y_2 and y_3; this assumption is required since we won't know what part of the plot belongs to the category estimated. That is the situation we currently have to live with. Generally violation of this assumption will result in higher estimates of accuracy than actually obtained.

Continuous variables will have to be put into classes in order to determine whether mapped correctly or not. This can be done objectively; for example, for percent tree cover use the 10 classes 0 to 10%, 10+ to 20%, ….90+ to 100%.

We then have the following determination for each of the *n* plots *k* ($k = 1,\ldots,n$):

If truth calls it y_{ij} and $y_{ij'}$ with plot area weights $w_{ij}^{y}, w_{ij'}^{y}$ such that $w_{ij}^{y} + w_{ij'}^{y} + w_{iother}^{y} = 1$ and x calls it the same with plot area weights $w_{ij}^{x}, w_{ij'}^{x}$ such that $w_{ij}^{x} + w_{ij'}^{x} + w_{iother}^{x} = 1$, then if $w_{ij}^{x} \le w_{ij}^{y}$, $w_{ij'}^{x} \le w_{ij'}^{y}$, correct classification for the plot gets a value of $(w_{ij}^{x} + w_{ij'}^{x})/n$ for p_0. $w_{ijother}^{y}$, $w_{ijother}^{x}$ are the percent (weights) of plot areas for which *y* or *x* or both define a condition on the plot not recognized by the other. The weight given to all partially or totally correctly classified plots is $(w_{ij}^{z1} + w_{ij'}^{z2})/n$ for p_{ij} where *z*1 and *z*2 are the smaller of w_{ij}^{y}, w_{ij}^{x} and $w_{ij'}^{y}, w_{ij'}^{x}$, respectively. The response variable for each plot *k* is then $p_{ij} = (w_{ij}^{z1} + w_{ij'}^{z2})/n$ where $0 \le p_{ij} \le 1/n$. Plots classified correctly or 0.80 correct are counted as 1/*n* and 0.80/*n*, respectively.

Calculate $p_0 = \sum_{i,j}^{k} p_{ij} w_{ij}$ and then compute the Kappa statistic in equation (66). By repeatedly taking n plots with replacement from the n sample plots *B* (say) 2,000 times and applying the above computation of the p_{ij} to each sample, we generate a series of *B* estimates for each cell of our contingency table as well as for producer and user accuracy and a Kappa statistic for each. With this bootstrapping we then can construct confidence limits around all the cells in the table as well as for the Kappa statistic by treating the *B* samples as independent estimates of the same quantities. An example of a contingency table with user, producer, overall accuracies, and the Kappa statistic based on results in Table 4 for a sample of *n* = 200 plots are:

User's accuracy for class 1 is 60.3/101 = 0.60,
for class 2 is 44.2/53 = 0.83, and
for class 3 is 30.8/46 = 0.70.

Producer's accuracy for class 1 is 60.3/70 = 0.86,
for class 2 is 44.2/73 = 0.61, and
for class 3 is 30.8/57 = 0.54,

with overall accuracy 135.3/200 = 0.68.

Then $p_0 = 0.68$ and $p_c = 0.33$ and $\hat{K} = \frac{0.68 - 0.33}{1 - 0.33} = \frac{0.35}{0.67} = 0.52$.

Then by repeatedly taking say 2,000 with replacement samples from the 200 sample plots, we compute for each sample the accuracies and the Kappa statistics again and construct confidence limits around the above producer's, user's, and overall accuracy as well as around the Kappa statistic.

The contingency tables are the basic product from the accuracy assessment. Users should study those tables in order to attempt to explain the causes of misclassifications. Some are obvious while others need investigating. Misclassifications may result from problems with the technology used, user errors, registration errors, errors in the final preparation of map products, or in calculations in the accuracy assessments. Studying the results is essential in that it may explain or uncover errors that can be corrected.

Table 4. A numerical example of a contingency table for forest cover class.

	Cover class 1	Cover class 2	Cover class 3	Row totals
Map class 1	60.3	20.7	20.0	101
Map class 2	2.6	44.2	6.2	53
Map class 3	7.1	8.1	30.8	46
Column totals	70	73	57	200

It is also desirable for a manager to know how serious a misapplication of a treatment to an area may be expected to be if the area is thought to belong in one category when in fact it belongs to another one. There would be different consequences in applying a treatment to a category close to the desired one than to a very different one.

Summary of what is needed for accuracy assessments:

- Define truth for the variables of interest and where and how to measure it. Minimize, or if possible, eliminate measurement errors in truth by observers.
- Decide on using either pixel accuracy assessments or polygon accuracy assessments.
- Ensure an adequate sample size in each of the categories of interest for the variables of interest.
- Define different types of accuracy or give some of these a different label than accuracy.
- Determine the implications of achieving a stated accuracy in terms of making correct or incorrect management decisions.
- Combine/integrate the accuracy assessments for the variables of interest and use the information also to improve the maps developed.

Global Positioning System for Spatial Location Needs

Spatial location is critical for success in forest inventory and monitoring because present needs require mapping this information.

Traditional methods of determining geographic location are still used, but more often than not, these former methods now supplement the Global Positioning System (GPS) when used in natural resources analysis. The GPS uses satellites to locate ground positions, usually within 150 meters, and often to less than 10 meters. With this system the location of aircraft and plots can be established rapidly with reasonably accuracy. GPS does not require the use of known geodetic markers for autonomous observation. Also, measurements can be made any time in any weather, with the exception of possibly large solar storms. However, because the measurements require a clear line of sight to the satellites, establishing location is difficult in a forest with a very dense canopy, deep valleys or gorges, or similar situations. Additionally, a number of errors do happen if the user does not use the GPS receiver correctly such as using the incorrect datum. Moreover, atmospheric and satellite-signal-path-errors occur naturally and need to be recognized and adjusted or reported inside the accuracy statements by the user. GPS has been used successfully as the basis of a sophisticated navigation and flight recording system controlling the acquisition of large-scale aerial photographs in West Australia (Biggs and others 1989, Biggs and Spencer 1990). Remote sensing has been oversold in the USA but will one day fulfill its promise partially because of reliable GPS systems. Accurate GPS will be even more reliable and precise in the future. For example, algorithms are being developed that correct satellite signals and make positioning more accurate. It behooves potential users to keep abreast of the technology because it is increasingly critical for inventory and monitoring.

Geographic Information System (GIS)

A Geographic Information System provides for entering, storing, manipulating, analyzing, and displaying spatial data. Data can be represented by points, lines, or polygons with the associated variables. Such data can be represented by raster or grid data on the one hand and by vector data on the other.

The raster system stores the data in a grid or pixel format with bounded geodetic values such as latitude and longitude, while the vector system uses a series of x, y coordinates to define the limits of the attribute of interest. Grid data are computationally easier to manipulate but usually require large amounts of storage space. Vector data require less storage and usually represent discrete data more accurately. Though the vector system may retain the shape of a discrete feature more correctly (has better resolution), it is computationally more time-consuming and difficult to render and analyze. Satellite imagery, digital pictures, and digital elevation models are examples of grid data; property boundaries, structure outlines, utility poles, and utility lines are examples of vector data. A GIS is a computerized system that can play a critical role in inventory, in manipulating and processing data, and in assessing land use and land cover. It has emerged as an effective tool in defining and focusing discussion relative to the merits of alternative land use allocations. For example, it gives the analysts the ability to simulate the effects of changes in management (Green 1993).

The GIS should have the capabilities to:

- Input many forms of data such as: analog and digital maps, textual or tabular information, and images.
- Store and maintain information with the necessary spatial relationships.
- Manipulate data, as in search and retrieval, and to do computations efficiently on the data.
- Provide levels of modeling taking into account data interrelationships and possible cause-and-effect responses of the relevant factors.
- Present tabulations, video displays, and computer generated maps of existing or derived information.

A good GIS depends primarily on good data. In addition high-speed computers, a variety of peripheral input-output devices, and powerful software are required. Articles by Congalton and Green (1992), Green (1993), and Bolstad and Smith (1992) give a good overview of GIS and Lachowski and others (1992) present a useful example of integrating remote sensing and GIS.

Small Area Estimation

There is considerable interest in management agencies to have reliable spatial information. In the past foresters and other land managers cruised or sketch mapped an area usually to decide what is where. Managers avoided statistical sampling because it might give reliable data on how much was there but not where. Frequent legal challenges changed this in the USA. Now interest is in obtaining reliable (defensible) mapped and statistical data together. One such area of current research is referred to as small area estimation, basically a model-building approach using statistical data in combination with ancillary data such as TM, GIS, topographic maps, and other related information.

Small area estimation techniques represent a substantial improvement in terms of quality of data, especially in defensibility of data-based management decisions relative to what used to be done when managers relied on subjective information. Small area estimates have been claimed to have standard errors similar to those for classical sampling. The trouble is the comparison is made for the entire population of interest whereas managers are also interested in predictions for much smaller areas such as polygons used as a basis for management. Standard errors for individual predictions can be large, as one would expect, given the variability encountered on the ground in forests.

For successful small area estimation, two conditions need to be met. First of all there should ideally be a good correlation between sampled and non-sampled areas either nearby or from other ancillary sources such as remote sensing. This usually requires a much more intensive grid than the 5,000-m grid now used by FIA. Secondly the spatial locations for both the sampled areas and the ancillary data need to be accurate. Given these conditions it should be possible to develop good prediction models.

Considerable work in small area estimation of forest resources is now being done. For example, there is innovative work being done in Finland where the more homogeneous conditions relative to other countries may make small area methods more useful. Multiple imputation methods (including regression models) and k-nearest neighbor techniques have been proposed for continuous variables. In these techniques, field sample information is extrapolated to the entire population where information on sample locations is input to non-sampled locations by some criteria such as similar

TM readings for the sampled and non-sampled locations. In multiple imputations for each unit without sample data, a series of say 100 predictions are made using randomly selected data and an underlying model and database. Then the data sets are analyzed separately and pooled into a final result, usually an average of the results.

Franco-Lopez (1999) reviews methods for projecting and propagating forest plot and stand information. As he notes, considerable effort has been expanded in Nordic countries combining forest monitoring information, remote sensing, and geographic information systems (GIS) to develop maps for forest variables such as cover type, stand density, and timber volume with emphasis on the k-nearest neighbor technique. He confides that while his results are poor for Minnesota, they are representative of those obtained by other methods in this region.

Lin (2003) presents a semi-parametric bootstrap method for estimating the precision of estimates. In general: 1. Fit the best fitting model say $y_i = \alpha + \beta x_i + \varepsilon_i$ resulting in the estimated model: $y_i = \hat{\alpha} + \hat{\beta} x_i$. 2. Compute the residuals $\hat{\varepsilon}_i$ and calculate the scaled residuals $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \sum_{i=1}^{n} \hat{\varepsilon}_i / n, i = 1, ..., n$. 3. Bootstrap the residuals $\tilde{\varepsilon}_i, i = 1, ..., n$, i.e. take a sample of n residuals with replacement from the n residuals. Do this say 1,000 times, each sample constituting a bootstrap sample. 4. For each bootstrap sample compute $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i + \tilde{\varepsilon}_i$. 5. Refit the model to each of the bootstrap samples using the sample points $\hat{y}_i, x_i$ and predict for each of the samples at the desired locations; the variability between these estimates is then used for the bootstrap variance for that location. Lin's results for predicting mortality, total basal area, and number of live trees on the Siuslaw National Forest in Oregon showed errors averaging nearly 100 percent for plots on a 1.36-km (0.85-mile) grid base using data from sampled 1-ha plots on a 2.72-km (1.7-mile) grid.

To obtain reliable predictions today, additional information is required, such as that available from improved remote sensors or large-scale photos and various maps combined with expertise from local ecologists. Also, at present it is still necessary to correct for location errors with models. Making such corrections requires considerable information on the extent and location of the errors. Hopefully, improvements in GPS-type sensors will reduce location errors in the future and improve results from small area estimation techniques.

VII. Sampling for Rare Events

Sampling rare populations is an order of magnitude more difficult than sampling common ones. Yet, assessing such populations can be critical. For example, the world is losing many plant and animal species and people may want to preserve at least some of these species. Knowing how many of a species exist and where and why in specific areas is critical for their preservation. The costs of locating rare populations are considerable and often exceed the costs of measurement. The fundamental problem is that several attempts may be needed to identify sample units with the rare trait in the overall population such as rare mushroom or tree species. Also, identifying the actual species or other attributes may require very specialized knowledge that only a few people have. Possible approaches are:

1. Screening. A large sample from the total population is examined to identify members of the rare population or at least areas where it is more likely to occur. If the latter is possible then such areas are sampled with much higher intensity than other areas for frequency of occurrence.
2. Multiplicity sampling and adaptive sampling. Basically, these are techniques that rely on locating some of the units with the rare attribute and then obtaining additional information about them, which is then used to locate others, thus reducing the cost of the survey.
   a. In multiplicity sampling a selected sample unit yields information about itself as well as about other units. Obviously this is more applicable to human surveys than vegetation surveys.
   b. In adaptive sampling a sample of units is selected probabilistically and units in the neighborhood of a sampled unit are added if the attributes of interest for those units satisfy a given criterion. The cleverness of the approach is that all units in the population are put in non-overlapping clusters and all units in the sample clusters are measured. Clusters can vary greatly in size and shape. Adaptive sampling is a probabilistic procedure but is hard to implement, and analysis of the results is difficult.
3. In multiple frame sampling, a sample is taken from an existing partial list and an additional one from the total population to screen for units with the characteristics of interest. The weakness of this approach is the overlap of frames (for which Kalton and Anderson, 1977 give some solutions) and the expense of screening and sampling the screened part of the population.
4. In snowball sampling, a necessary condition is that units contain information about each other. Then a frame of units is created in the rare population by sampling a few units and through them identifying others. Clearly again, this is more likely to be fruitful with human populations than with vegetation. Once a frame has been developed, a probabilistic sample is drawn. The weakness of this approach is the degree of completeness of the frame. An advantage is that rare units are identified more quickly than with other methods.
5. Sequential sampling. Select an initial probabilistic sample of sufficient size to give the desired sample size (*n*) of members of the rare population based on the rate of incidence observed. This will yield n_1 members of the rare population and an estimate of incidence. If $n_1 < n$, a second sample is selected to produce the remaining $n - n_1$ members of the rare population based on the incidence obtained in the first sample. This procedure is generally expensive and hence is not practical in most vegetation surveys.

Several of these techniques may become more useful in vegetation and animal population surveys as rapid, simplified DNA identification methods are developed.

VIII. Multiple Level Sampling

Previously, we discussed single-level surveys where either the variables of interest or these variables plus covariates on the same sample units were measured. When covariates were measured, we assumed that the covariate values for all units in the population or, at least the population total, were known. Often the covariates are useful for estimation and are cheaper to collect but unknown. It often pays to collect covariate information on a large sample and the variable of interest on a subsample.

This approach is referred to as multilevel sampling. For example, in a timber sale we might obtain ocular estimates of diameter at breast height D (and hence D^2) for a large sample of trees and measure actual volume of some of these trees. Or, in estimating recreational use of an area, we may use traffic counters recording counts of vehicles at the entrance to the park on a large number of days and actually count the number of users on a subset of these days.

Multilevel sampling can be separated into multiphase and multistage sampling.

Multistage Sampling

Refers to sampling designs where the ultimate sample units, called elements, are selected in stages. Samples at each stage are taken from the sample units comprising clusters of units selected in the previous stage. Interest is in estimation of attribute totals or means per element, such as biomass per tree rather than per ha. The population is first divided into a number of primary sample units (PSU), some of which are selected as the first stage sample. These selected PSUs are then subdivided into a series of secondary sample units (SSU), some of which are randomly selected as the second stage sample. This process can be repeated of course with additional stages if necessary. The procedure has the advantage of concentrating work on a relatively small number of PSUs after which much less effort is usually needed to obtain the second and later samples.

The main reasons for selecting a multistage sample are:

1. Drawing a set of units from a population such as trees in a large forest or recreation users of a park over a full season is expensive. It is difficult to obtain a list of all the trees and even more difficult to determine all users of a park.
2. Even if a list of population units was available, efficiency might dictate that groups of units (clusters) rather than single units be chosen and that only some units in each cluster are measured. For example, it is usually cheaper to sample 20 randomly located clusters of 30 trees in a forest than 600 randomly located trees and we may want to only sample 10 out of the 30 trees in each cluster because of the homogeneity in the cluster or the expense of measuring all 30 trees. In sampling recreation users, it is clearly easier to select and subsample random days on which to interview all users rather than attempt to randomly sample individual users or days, respectively.

Generally, though as indicated earlier, there is a definite tradeoff in efficiency between cluster sampling and random sampling of units because units close together are often more similar than those further apart and it often pays to measure only some of them in each selected cluster.

Sampling can be in a large number of stages. We illustrate how this works with the simple and often practically useful situation of 2-stage sampling with SRS at each stage. Assume N groups or clusters with M_i units (i = 1, …, N) in the i^{th} cluster. Our total of interest can now be written as:

$$Y = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} Y_{i.} \qquad (67)$$

In 2-stage sampling a random sample of n is selected out of the N clusters but instead of measuring all units in the cluster, a random sample of m_i units is chosen in each. Thus, the cluster total $Y_{i.}$ is first estimated by

$$\hat{Y}_{i.} = M_i \sum_{j=1}^{m_i} \frac{y_{ij}}{m_i} \tag{68}$$

for each of the *n* clusters sampled. Our estimated total is:

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} M_i \left(\sum_{j=1}^{m_i} \frac{y_{ij}}{m_i} \right) = \frac{N}{n}\sum_{i=1}^{n} \hat{Y}_{i.} \tag{69}$$

with variance

$$V(\hat{Y}) = \frac{N^2 M_a^2 (1-f)\sigma_b^2}{n} + \frac{N}{n}\sum_{i=1}^{N} \frac{M_i^2 (1-f_i)\sigma_{wi}^2}{m_i} \tag{70}$$

where:

$$\sigma_b^2 = \frac{\sum_{i=1}^{N} \left(\frac{M_i}{M_a} \bar{Y}_i - \bar{Y} \right)^2}{(N-1)}\text{, the between-cluster variance, and}$$

$$\sigma_{wi}^2 = \frac{\sum_{j=1}^{M_i} \left(y_{ij} - \bar{Y}_{i.} \right)^2}{M_i - 1}\text{, the within-cluster variance.}$$

Here $f = \frac{n}{N}$, $f_i = \frac{m_i}{M_i}$, $\bar{Y}_{i.} = \frac{Y_{i.}}{M_i}$, $\bar{Y} = \sum_{i=1}^{N} \frac{Y_{i.}}{N}$, and $M_a = \sum_{i=1}^{N} \frac{M_i}{N}$, the average number of SSUs per PSU. Similarly, an unbiased variance estimator $v(\hat{Y})$ is:

$$v(\hat{Y}) = N^2 M_a^2 (1-f)\frac{s_b^2}{n} + \frac{N}{n}\sum_{i=1}^{n} \frac{M_i^2 (1-f_i) s_{wi}^2}{m_i} \tag{71}$$

where:

$$s_b^2 = \frac{\sum_{i=1}^{n} \left(M_i \bar{y}_i - \bar{y} \right)^2}{n-1}, \quad \bar{y}_{i.} = \sum_{j=1}^{m_i} \frac{y_{ij}}{m_i}, \quad \bar{y} = \frac{\hat{Y}}{NM_a} \text{ and } s_{wi}^2 = \sum_{j=1}^{m_i} \frac{(y_{ij} - \bar{y}_{i.})^2}{m_i - 1}.$$

There is a considerable literature on multistage sampling, but this subject is still best discussed in the book by Murthy (1967).

Multiphase Sampling

In multiphase sampling the same size of sample units are retained at each level (phase) but with fewer sample units selected at each consecutive one. In the last phase the variable of interest is measured and is combined with covariate information from the early phases either in design (stratification or pps sampling) or estimation (regression or ratio estimation). In multiphase sampling a complete frame of units is required since a sample of units is selected at each phase. The main reason for using multiphase sampling is to reduce the cost of sampling by collecting a large amount of relatively cheap information on covariates that are correlated with the variables of interest and then measuring the variables of interest on a smaller sample. Stratified double sampling and double sampling for regression or ratio estimation are two examples. Specifically:

1. For stratified double sampling, the large (first phase) sample information is used to construct strata from which the second phase samples are selected. Typically this is done if interest is in specific subpopulations (strata) or the strata are more homogeneous than the overall population so that efficiency is gained by stratification. For example, in traditional large-scale timber surveys we might have a large sample of say $n^{'}$ 1-ha plots from remote sensing or photos classified into primarily large timber, pole timber, and regeneration. Clearly, if interest is in volume, those three strata would be of interest in their own right and are likely to be much more homogeneous (if sufficiently well done by remote sampling) than the overall population. A subsample of those 1-ha plots would then be sampled on the ground for volume by stratum. Similarly in sampling a large park for recreation use, we might take a large sample of photos on sample days to count users, use that information to divide the park into strata of heavy, moderate, and low use days, and then sample these three strata on a subset of those same sample days. The estimator of the total in both cases is:

$$\hat{Y}_{dst} = N\sum_{h=1}^{K} w_h \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} = \sum_{h=1}^{K} \hat{N}_h \bar{y}_h = N\bar{y}_{st} \tag{72}$$

where K = number of strata, $\hat{N}_h = \frac{Nn_h^{'}}{n^{'}}$ is the estimated number of sample units in stratum ***h*****,** $w_h = \frac{n_h^{'}}{n'}$ is the estimated weight for stratum *h* for the first phase sample with $n_h^{'}$ and n' the first phase sample sizes for stratum h and overall respectively, n_h is the second phase sample in stratum *h*, and $\bar{y}_{st}$ is the estimated mean for stratum *h* for the sample of n_h units in that stratum. The variance of this estimator is:

$$V\left(\hat{Y}_{dst}\right) = N^2S^2\left(\frac{1}{n'} - \frac{1}{N}\right) + N^2\sum_{h=1}^{K}\frac{W_hS_h^2}{n'}\left(\frac{1}{v_h} - 1\right) \tag{73}$$

with S^2 the population variance of y, S_h^2 the variance of *y* in stratum *h*, $v_h = n_h / n \le 1$ and $\bar{y}_h$ and $\bar{y}_{st}$ the sample mean for stratum *h* and overall sample mean for stratified sampling respectively.

An almost unbiased sample estimator of $V(\hat{Y}_{dst})$, if both 1/*N* and $\frac{1}{n'}$ are negligible, is:

$$v\left(\hat{Y}_{dst}\right) = N^2\sum_{h=1}^{K}\frac{w_h^2 s_h^2}{n_h} - N\sum_{h=1}^{K} w_h s_h^2 + N^2\frac{g_1}{n'}\sum_{h=1}^{K} w_h(\bar{y}_h - \bar{y}_{st})^2 \tag{74}$$

where $$g_1 = \frac{N - n'}{N - 1}.$$

Strata may be of different degrees of interest and vary in homogeneity, so varying sampling rates may be desirable. This requires knowledge of or an estimate of the variability within the strata in order to allocate *n*. If such knowledge is available, one can then optimally allocate the sample to the strata. Assume that there is information available or easily collectable on a variable *x* correlated with y. Then applying the simple cost function:

$$C = C'n' + n'\sum_{h=1}^{K} C_h n_h$$

where C' is the cost of classifying a unit for the first phase and C_h is the cost of measuring a unit in stratum *h*, the expected cost *E(C)* is:

$$E(C) = C'n' + n'\sum_{h=1}^{K} C_h v_h W_h \,. \qquad (75)$$

Then the optimal n' can be computed from by substituting

$$\hat{v}_h = s_{yh}\sqrt{\frac{C'}{C_h}\left(s_y^2 - \sum_{h=1}^{K} w_{0h} s_{yh}^2\right)}$$

for v_h where s_y^2 and s_{yh}^2 are the estimated variance for variable *y* in the population and stratum *h* respectively and w_{0h} is the estimated stratum weight for stratum *h* based on the preliminary information. More complex cost functions are discussed in the literature, especially Hansen and others (1953), but usually insufficient information is available to assume a better cost function, so it makes sample size computations more difficult and sample size determination seems fairly insensitive to improved cost functions.

2. For double sampling with ratio or regression estimators, a linear relationship is assumed between the covariates and the variables of interest as shown in the general linear model in (39).

For instance in the timber example above, one may have confidence that the information on the 1-ha remotely sensed or photo plots is linearly related to the same information as measured on the ground. Or, similarly, the photo counts of recreational users might be linearly related to the actual counts on the ground. Clearly, whether such a linear relationship exists as a useful approximation or there is a useful but unknown relationship between the remote sensing and the ground information in both cases determines whether stratified double sampling or double sampling with ratio/regression estimation is more efficient and reliable. The regression estimator of the overall total is:

$$\hat{Y}_{gr} = \sum_{i=1}^{n}\frac{y_i}{\pi_i} + a_{gr}\left(\hat{N}_1 - \sum_{i=1}^{n}\frac{1}{\pi_i}\right) + b_{gr}\left(\hat{X}_1 - \sum_{i=1}^{n}\frac{x_i}{\pi_i}\right) \qquad (76)$$

where:

$$a_{gr} = \frac{\sum_{i=1}^{n}\frac{y_i}{\pi_i v_i} - b_{gr}\sum_{i=1}^{n}\frac{x_i}{\pi_i v_i}}{\sum_{i=1}^{n}\frac{1}{\pi_i v_i}}$$

$$b_{gr} = \frac{\sum_{i=1}^{n}\frac{1}{\pi_i v_i}\sum_{i=1}^{n}\frac{x_i y_i}{\pi_i v_i} - \sum_{i=1}^{n}\frac{y_i}{v_i \pi_i}\sum_{i=1}^{n}\frac{x_i}{v_i \pi_i}}{\sum_{i=1}^{n}\frac{1}{\pi_i v_i}\sum_{i=1}^{n}\frac{x_i^2}{v_i \pi_i} - \left(\sum_{i=1}^{n}\frac{x_i}{v_i \pi_i}\right)^2}$$

with $\hat{N}_1 = \sum_{j=1}^{n'}\frac{1}{\pi_{ja}}, \quad \hat{N}_2 = \sum_{i=1}^{n}\frac{1}{\pi_i}, \quad \tilde{N}_s = \sum_{i=1}^{n}\frac{1}{\pi_i v_i}, \quad \hat{X}_1 = \sum_{i=1}^{n'}\frac{x_i}{\pi_{ja}}, \quad \hat{X}_2 = \sum_{i=1}^{n}\frac{x_i}{\pi_i}, \quad \tilde{x}_s = \frac{\sum_{i=1}^{n}\frac{x_i}{\pi_i v_i}}{\tilde{N}_s}$

where π_{ja} is the probability of selecting unit j in the sample of n' units and π_i the probability of selecting unit i in the sample of n units and $\tilde{y}_s = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i v_i}}{\tilde{N}_s}$. The π_i may not always be computable.

Deriving a classical variance estimator for this estimator is difficult and this is an example where bootstrap variance estimation would be the method of choice.

Illustration: A large sample of *n'* plots is measured for plot basal areas on aerial photos. These could be stratified into *K* strata, selecting either a subsample of *n* plots in the *K* strata or a SRS of *n* out of the *n'* plots which are then measured on the ground**.** Using BAT_i and VT_i to denote basal area on plot *i* as measured on the photo plots and volume as measured on the ground plots,

we then have:
n' plots with $BAT_i, i = 1,...,n'$
n plots with $BAT_{hi}, h = 1,...,K, i = 1,...,n_h$ and $VT_{hi}, h = 1,...,K, i = 1,...,n_h$ for stratified double sampling or $BAT_i, i = 1,...,n$ and $VT_i, i = 1,...,n$ for double sampling with regression.

Whether stratified double sampling or double sampling with a regression estimator would be used, depends on the relationship expected between BAT_i and VT_i. If there is expected to be a linear relationship, regression estimation would be used, otherwise double sampling for stratification is indicated.

For stratified double sampling one would use (72) and (74) to estimate total volume and its variance. For double sampling with regression one would use (76) with a bootstrap variance estimator.

If one expects the relationships between the covariates and the variables of interest to go through the origin approximately, a double sampling with a ratio of means estimator can be used:

$$\hat{Y}_{drm} = \left(\frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{x_i}{\pi_i}} \right) \hat{X}_1 = \left(\frac{\hat{Y}_{HT}}{\hat{X}_{HT}} \right) \hat{X}_1 . \tag{77}$$

In this case too it is best to use bootstrapping to estimate the variance of $\hat{Y}_{drm}$. Here too the π_i may not always be computable.

Multilevel sampling methods in forestry are common especially for large scale surveys. For example:

1. Double sampling for stratification is used in large-scale surveys such as FIA. Areas are stratified usually into forested vs. non-forested areas by either photography or more commonly now by data collected from remote sensing sources such as the Landsat Thematic Mapper Satellite (TM) and then ground plots are measured in those strata. In the past, with primary interest in timber, prestratification was used. Now post-stratification is used because plots are grid-based. Newer remote sensing sources will define small features on the ground better and locations of both the ground and remote sensing information can be pinpointed more accurately with improved GPS units. It is likely that more detailed stratification and regression estimation will improve estimation in the future.
2. VRP sampling with subsequent selection of trees by either Poisson sampling proportional to estimated tree heights or another subsampling scheme were frequently used in timber sales.

Clearly combinations of multiphase and multistage sampling can be desirable too. For instance, in example 1 above we might select a random sample of trees on the selected ground plots. This design then would be double sampling for stratification with random subsampling.

IX. Monitoring Over Time

Managers of biological resources are always interested in change over time in timber volume, mortality, wildlife habitat, degree of urbanization, change from forest land to agricultural land, etc.

There are three major sampling options to consider in sampling over time:

1. Complete replacement sampling (CRP)
2. Complete remeasurement sampling (CR)
3. Sampling with partial replacement (SPR), a combination of a and b.

In CRP sampling, a completely new set of sample plots is used each time. Such a design is simple and cheap to implement since plot locations do not have to be monumented for future use and one does not have to worry about the plots being treated different from other parts of the population or changes in the underlying population. CRP sampling is efficient for estimating current attributes but not efficient for estimating change relative to CR and SPR.

In CR sampling all sample plots are remeasured periodically. This requires that they remain representative of the population over time, so that the plots should not be visited excessively, and should be treated no differently from other parts of the population. CR sampling is the most efficient of the methods available for change estimation.

In SPR a random subset of the permanent plots is remeasured as well as a new set of plots, i.e., it is a combination of CRP and CR sampling. Regression estimation between the remeasured and new plots is used to “update” the plots that were not remeasured. SPR can be efficient when trying to balance precision between current and change estimation.

Duncan and Kalton (1987) summarized the properties of the three options nicely. They also list another method that is a combination of the other three (Table 5).

Both CR and CRP sampling are special cases of SPR sampling from an estimation point of view so we only present SPR sampling for two occasions here:

If n sample units are selected from N units at both occasions with m units common to both, then $u = n–m$ units are not shared.

Let $\hat{Y}_{im}, \hat{Y}_{iu}$, and $\hat{Y}_{in}$ equal the estimates of Y_i, the population total on the i^{th} occasion (i = 1,2), based on the m, u, and n units respectively, $\hat{\beta} =$ the regression coefficient estimator based on the m common units, σ_1^2 and σ_2^2 the variances of y at times 1 and 2, σ_{12} the covariance of y between times 1 and 2 and ρ the correlation between measurements at times 1 and 2. Then an unbiased estimator of Y_2 based on the u new units at time 2 is:

$$\hat{Y}_{2u} = N \frac{\sum_{i=1}^{u} y_{2i}}{u}, \tag{78}$$

with variance:

$$V(\hat{Y}_{2u}) = N^2 \sigma_2^2 / u, \tag{79}$$

and variance estimator:

$$v(\hat{Y}_{2u}) = s_2^2(1)/u \tag{80}$$

with $s_2^2(1)$ the within-sample variance of the u y_{2i} measurements.

Equivalently, a regression-based estimator of Y_2 using the m common units at times 1 and 2 to update the total from time 1 is:

$$\hat{Y}_{2mr} = \hat{Y}_{2m} + \hat{\beta}(\hat{Y}_{in} - \hat{Y}_{im}) \tag{81}$$

Table 5. Objectives and properties of four remeasurement designs (adapted from Duncan and Kalton 1987).

Sampling objective	Complete replacement sampling (CRP)	Complete remeasurement sampling (CR)	Sampling with partial replacement (SPR)	Combination of CR with CRP or SPR
a) Estimate population parameters at distinct times. **b)** Estimate average values of population parameters.	Automatically takes into account population changes.	Needs mechanism for taking population changes into account.	Needs mechanism for taking into account population changes during life of replacement group. Composite estimates can be used to produce efficient estimates.	Remeasurement component needs mechanism for taking population changes into account.
c) Estimate net change.	Estimates combined effect of changing values and changing population.	Needs mechanism for taking population changes into account. Variance of change reduced by positive correlation of values between surveys.	Needs mechanism for taking into account population changes during life of remeasurement. Composite estimation can be used to produce efficient estimates.	Remeasurement component needs mechanism for taking population changes into account. Variance of change in remeasurement component reducedby positive correlation of values between surveys.
d) Estimate components of change over time. **e)** Aggregate data for individuals over time.	Not possible.	Well-suited for these populations.	Can be used for change estimation or aggregate information over time periods shorter than the time a sample is to be replaced in sample. Only the sample to be replaced can be used.	Complete remeasurement component is well-suited for these purposes. Not possible for complete replacement surveys component.
f) Collect data on events occurring in specific time periods.	Not possible.	Can construct long-term history of events by combining data from several surveys.	Can construct long-term history of events but on a more limited basis than complete remeasurement surveys.	Can construct limited long-term history of events.
g) Cumulate samples over time.	Excellent for static characteristics and for new events.	Not useful for static characteristics, but useful for new events.	Of some use for static characteristic and useful for new events.	Complete remeasurement survey component is excellent. Complete remeasurement survey component is useful for new events but not for static characteristics.

with variance:

$$V(\hat{Y}_{2mr}) = N^2\left(\frac{\sigma_2^2}{m}\right)\left(1-\frac{u\rho^2}{n}\right). \qquad (82)$$

Combining estimates $\hat{Y}_{2u}$ and $\hat{Y}_{2mr}$ to obtain an improved estimate of Y_2 is usually done by weighing them inversely proportional to their sample variances, so we obtain:

$$\hat{Y}_2 = \frac{\left(\hat{w}_1\hat{Y}_{2u} + \hat{w}_2\hat{Y}_{2mr}\right)}{\hat{w}} \qquad (83)$$

with approximate variance $V(\hat{Y}_2) \doteq 1/w$

with $w = w_1 + w_2 = 1/\sigma_1^2 + 1/\sigma_2^2$

and variance estimator: $$v(\hat{Y}_2) = \left[1 + \frac{4}{\hat{w}^2}\sum_{i=1}^{2}\frac{\hat{w}_i(\hat{w}-\hat{w}_i)}{d_i}\right]\hat{w} \qquad (84)$$

with $\hat{w}_1 = \frac{1}{v(\hat{Y}_{2u})}$, $\hat{w}_2 = \frac{1}{v(\hat{Y}_{2mr})}$, $\hat{w} = \hat{w}_1 + \hat{w}_2$, $d_1 = m-1$, $d_2 = u-1$, and w is estimated by $\hat{w}$. Two estimators of change $\triangle Y$ are possible, the most obvious one being:

$$\triangle\hat{Y}_1 = \hat{Y}_2 - \hat{Y}_1. \qquad (85)$$

A desirable property of such an estimate is that it is consistent with the estimates at the two occasions.

A more efficient estimator than $\triangle\hat{Y}_1$ in general takes advantage of the regression based on the *m* common units as in $\hat{Y}_2$ above. This estimator is:

$$\triangle\hat{Y}_2 = \frac{\hat{w}_1\triangle\hat{Y}_1 + \hat{w}_2\triangle\hat{Y}_2}{\hat{w}}. \qquad (86)$$

Here $\triangle\hat{Y}_1$ and $\triangle\hat{Y}_2$ are the estimators of change from the *m* remeasured and *u* unmatched plots respectively, where

$$\hat{w}_1 = \frac{s_1^2}{m} + \frac{s_2^2}{m} - 2\frac{s_{12}}{m}, \quad \hat{w}_2 = \frac{s_1^2}{u} + \frac{s_2^2}{u}$$

are unbiased estimators respectively of:

$$w_1 = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{m} - 2\frac{\sigma_{12}}{m}, \quad w_2 = \frac{\sigma_1^2}{u} + \frac{\sigma_2^2}{u},$$

$\hat{w} = \hat{w}_1 + \hat{w}_2$ estimates $w = w_1 + w_2$, and

$$V(\triangle\hat{Y}_2) = 1/w \qquad (87)$$

with approximate variance estimator:

where: $$v\left(\triangle\hat{Y}_2\right) = \left[1 + \frac{4}{\hat{w}^2}\sum_{i=1}^{2}\frac{\hat{w}_i(\hat{w}-\hat{w}_i)}{d_i}\right]\hat{w} \qquad (88)$$

$$d_1 = m-1, d_2 = u-1.$$

Problem: Show how CR and CPR sampling are special cases of SPR sampling.

Answer: Set $\rho = 0$ in (82) to get the variance for CR sampling and $\rho = 1$ to get the variance for CPR sampling.

The matching proportion in SPR sampling depends on the correlation ρ between the measurements at the two times. It should not exceed 0.5 for optimizing the estimator $\hat{Y}_2$. SPR sampling quickly becomes much more complicated in estimation when more than two occasions are measured (Schreuder and others 1993). But a serious disadvantage of SPR sampling, that of variance estimation, has been eliminated. With bootstrapping it should be simple to generate variance estimates for any number of remeasurements and estimation schemes.

All of the FIA units in the USA now use complete remeasurement sampling, although one formerly used sampling with partial replacement. In general, SPR sampling is probably most efficient but becomes quite complex from an analysis point of view, which makes it hard to deal with the numerous special requests of estimates for specific subpopulations of the survey population covered.

X. Building Models and Cause-Effect

The statement by Box and Draper (1987) that "all models are wrong; the practical question is how wrong do they have to be to not be useful" is generally accepted in the statistical world, and can be paraphrased as "all models are wrong and some are useful." The utility of models is often assessed by the degree of correlation between the variables of interest and covariates, but note that correlation does not prove causation (Kish 1967).

Much if not all of research revolves around model building, and the potential misuse of models has been greatly facilitated by the ready availability of computers and easy use of regression programs. Ideally, a researcher observes the real world or carefully studies substantive scientific theories. Models are then developed on the basis of the insights accorded, recognizing the fact that besides the explanatory variables, there are other sources of variation to be considered. Kish (1967) separates all sources of variation into four classes:

1. The explanatory or experimental variables that are the objectives of the research in explaining or establishing a relationship between both the dependent (often called the response variables in this context) and the independent (often called predictor variables in this context) variables.
2. Extraneous variables that can be controlled either in sample selection or estimation.
3. Extraneous (unmeasured, often unmeasurable) variables that may be confounded with the variables in class 1 above.
4. Extraneous, difficult to control or uncontrollable variables that have to be treated as randomized errors. In ideal experiments, they can be randomized, whereas in surveys they can only be assumed to be randomized.

In all research one wants to place as many extraneous variables as possible in class 2. Since this usually cannot be done, we have experiments and surveys. Experiments, the conduct of a systematic, controlled test or investigation, tries to control the variables in class 3 as much as possible by trying to place all of the third class of variation into the fourth through randomization. In an ideal experiment, there are no variables in the third class. In an ideal survey, all variables in class 3 are separated from those in class 1 through regression adjustments, matching of units, and standardization.

If there was complete command over the research situation, one could introduce the desired effects and measurements into controlled and randomized portions of the target population with firm experimental controls and build a "true" model (Kish 1967). Such situations are rare so that we have experiments that are strong on control through randomization but weak on representing actual populations of interest and frequently weak on the natural setting of the model being built. Surveys often are feasible when experiments cannot be done, the most obvious being that we do not experiment on humans. Surveys are strong on representation but are often weak on control. The latter is a ready explanation of why so many studies are published claiming this or that chemical is bad for you and subsequent research does not support such claims.

Often survey data are used to build models, to lead to a better understanding of what is going on. Many models appear to have poor predictive ability; for example, this is true for the staple of forest research: building growth and mortality models. A missing ingredient is key data that would help in identifying cause-effect relationships such as daily rainfall, atmospheric deposition, soil moisture content, etc. Such data cannot be collected yet in a practical manner in conjunction with natural resources surveys, but development of new instrumentation should make that feasible some day. Until this information is routinely available for the plots, prediction models for growth, mortality, erosion, and other key variables are unlikely to be reliable.

Large-scale surveys such as FIA and the Natural Resources Inventory (NRI) of the National Resources Conservation Service (NRCS) can establish trends in change for large areas, can be used to suggest and identify potential cause-effect relationships, and can suggest useful hypotheses to document relationships (Olsen and Schreuder 1997). Inferences about possible cause-effect relationships have to be interpreted cautiously because screening of data makes it difficult to define the population of inference (see for example Schreuder and Thomas 1991). It is unfortunate in the USA that there are two natural resources surveys, Forest Inventory and Analysis of the USFS and the NRI of the NRCS where both agencies are in the USDA. Narrowly focused surveys seem to be the rule in most other countries too as evidenced from descriptions in European Commission (1997). Complementariness of the data collected would make it more likely to identify promising cause-effect relationships for a wider range of resource variables. For example, it would be desirable to have the reliable information on soils collected by the NRI also available on the FIA plots to develop better growth and yield models.

Mosteller and Tukey (1977) identify three criteria, of which two have to be satisfied to infer cause-effect relationships: consistency, responsiveness, and mechanism. Consistency implies the presence and magnitude of the effect y, associated with a minimal level of the suspected causal agent x. Responsiveness is established by an experimental exposure of the suspected causal agent and reproducing the symptoms. Mechanism demonstrates the biological or ecological process causing the observed effect. To establish all three criteria is difficult. For example, the consistency between smoking and cancer was established in the 1950s. The responsiveness was well documented then and in the 1960s, but the actual causal mechanism was not established until the 1990s (Peifer 1997). And this linkage was relatively easy to establish because the effect is dramatic; see for example Taubes (1995).

Feinstein (1988) advocated the following scientific principles for establishing cause-effect: stipulate a hypothesis prior to analysis, study a well-defined cohort having a statistical factor in common, collect high-quality data, study possible explanations, and avoid detection bias.

Hill (1965) in the epidemiological literature suggests a weight of evidence approach consisting of nine criteria for inferring causality: strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experimental evidence, and analogy. Strength refers to having a high magnitude of an effect associated with exposure to the stressor; consistency to repeatedly observing the association of the observed effect and stressor under different conditions; specificity to the degree of the effect being more likely to be diagnostic of the stressor and the ease of associating it with an effect; temporality to the fact that the stressor always precedes the effect in time; biological gradient to the change in effect with corresponding changes in the stressor; plausibility that the association between effect and stressor is consistent with physical, chemical, and biological principles; experimental evidence that changes in effect are documented after experimental manipulation or through recovery of the population following relief of the stressor; and analogy is having similar stressors associated with similar effects. The more of these criteria that are satisfied, the more weight can be given to the evidence that there is probable cause.

Survey data can only provide information for identifying possible cause-effect relationships. Establishing that there is a correlation between possible cause and effect variables is a useful first step in this indentification For readers who want more information on how to maximize the possibilities of such identification, we refer them to Olsen and Schreuder (1997) and Gadbury and Schreuder (2003).

XI. Forest Sampling Situations

Pitfalls

Now that you have studied sampling intensively you may think you know what to do. This section covers some major errors committed by serious samplers over time in the USA.

1. Moving subplots**.** FIA used to select 0.4 ha (1 acre) primary sample units subsampled by a series of 5 or 10 VRP subplots. As is logical, in some cases some of the subplots might fall in a different condition than the center subplot. For example, subplots 1-3 might be sampling a pine plantation and subplots 4-5 in a hardwood stand. The decision was made as early as the 1930s and continued into the 1990s by several FIA units to keep all subplots in the same forest type. For example if subplot 1 (the center one) was in a pine plantation, all subplots not falling in the pine plantation would be moved by some systematic rule into the pine plantation. This procedure biases the results (Williams and others 1996).
2. Averaging conditions. Related to the above, another unit did not move the subplots in the above situation but made the equally undesirable decision to average forest types, i.e., they did not keep track of what type was being sampled. They would call the plot described in (1) above a mixed pine-hardwood stand.

   Situations 1 and 2 led to the interesting situation where two states in the USA that are quite similar showed huge differences in the area in pine/hardwood stands as a percentage of the total in forest.
3. VRP sampling to get 6-8 trees per point. A reasonable recommendation was made in a forest mensuration textbook that in VRP sampling one should select a prism factor yielding on average 6-8 trees/point. This recommendation was followed up incorrectly in several places in the Western USA. Field crews would take various prisms to the field with them and then select one at a location that would yield them between 6 and 8 trees. This biased approach surprisingly was supported by several prominent biometricians. In several “experiments” of this method, little or no bias resulted but one author got involved in a situation in California where such serious bias was noted (see Wensel and others 1980 and Schreuder and others 1981).
4. Misuse of model predictions. A FIA unit developed growth and mortality models based on growth and yield studies and used those models to update the information on plots they could not remeasure from a cost point of view. The predicted plot values were then used as real plot data for generating state-wide estimates.
5. Dropping subplots to meet production targets. FIA program managers put heavy emphasis on meeting production targets. This is why one unit approved of the elimination of subplot 4 to meet the production of 8 plots per 2 weeks for crews if they felt they could not meet their production targets. This biases the results, especially if crews decide to judiciously drop subplots 4 such as in difficult sampling conditions.
6. Forgetting probabilities of selection. A government agency selected a timber cruising sample using stratified sampling to obtain volume estimates for different strata. Ten years later they decided they wanted to revisit the locations for other purposes but had not kept track of the probabilities of selection. They wanted to treat the existing sample as a SRS for remeasurement purposes. Schreuder and Alegria (1995) illustrate how this may seriously bias results.
7. Treating subplots as plots because their information is considered uncorrelated. Clearly the subplots are not independent observations and hence should not be treated as such.
8. Different results by different agencies. Two agencies in the same department estimated vastly different areas of forest in several states. It turned out that this was due to differences in interpreting a common definition of forest, definition of what is a tree, and standards in measurement and estimation techniques. Several of these differences are also tied into

considering forest as a use class (Goebel and others 1998). A forester may prefer to see as much forest in a state as possible, while a range manager may see the same land as range.

Problem: Using a tree growth model developed from growth and yield study data, how would the predictions for such a model compare to the actual growth of trees of the same species on inventory plots?

Answer: Growth and yield studies typically use plots with 100 percent stocking levels with a much more favorable environment than that of inventory plots which are more likely impacted by insects and diseases, by human activities, etc. It is therefore more likely that the predictions will yield overestimates of the actual growth of the inventory trees.

Problem: An unmotivated crew using the situation described in #5 decides to always drop subplot 4 when they know it is difficult to measure. What are the consequences?

Answer: Clearly this will bias estimates for the area inventoried because it changes the probabilities of selection of the subplots and hence the plot that they are part of. It is really not possible to answer which way the bias will go. Some of the subplots may be in areas difficult to measure because they are in a swamp while others may be in highly productive areas where vines and underbrush make access to the subplots difficult and others may be on a very unproductive, steep cliff.

Problem: A well-intentioned crew using the situation described in #5 decides to measure subplot 4 only when they know it contains nice "timber" trees. What are the consequences?

Answer: This should bias certain estimates upward such as those for timber volume. It might not have much effect on estimates for variables not much correlated with niceness of trees such as number of trees or mortality.

Problem: A crew used multiple prisms selecting the one giving 6-8 trees at each point in VRP sampling. Generating estimates for the 100 VRP plots they took generates an estimate of 100,000 m^3 for the area. They sell the timber based on that amount of volume. The company that buys the timber finds only 60,000 m^3. They are not happy and file a lawsuit against the company that did the inventory. Both sides approach you, a reknown inventory expert, to testify on their behalf. Since truth is more important to you than money, you can pick either side. Which one would you pick?

Answer: It would be more sensible to take the side of the company that bought the timber. Certainly the inventory method used was faulty as indicated in #3 above.

Suggestions

Our experience leads us to believe that one can do better than what is available now in regards to surveys. It is hard to change existing surveys. Hence:

1. Be more flexible, less hidebound.
2. Do research on what is available and what can be done better.
3. Document well what you are doing.
4. Observe the "Keep it simple" (KIS) principle in design, less so in estimation.
5. Use creative, competent analysts.

6. Focus on the objectives. For example, there may be need to resolve potential conflicts between what is wanted for timber surveys for which there is often strong political support vs. ecology parameters where such support may be less powerful. Anticipate what may be needed in the future too.
7. Keep up on the world literature and contribute to it.
8. Define measurable variables (see Schreuder and others 1993, p. 292, specifically the warning by Innes for example).

Over time, we have learned the following lessons:

1. The objectives of a successful survey will change over time and will become more encompassing.
2. Don't lock yourself into existing approaches. Allow for change. An example is plot design where in the USA we have gone from rectangular plot sampling to variable radius plot or VRP sampling, to circular plot sampling and more than likely at a future date should move to sampling using different plots for different variables including a long rectangular or square plot closely tied to remotely sensed information. Over time, more information can be collected by remote sensing. Large-scale surveys are getting away from a pure timber orientation to being more ecologically based so that we are interested also in linear features such as ripararian areas, understory vegetation, and rare and endangered plant species. Because plots can be more accurately colocated on both the remote sensing and the actual ground plots by the use of geographic positioning systems (GPS) over time, and because more and more detail can be discerned with newer remote sensing platforms, more efficient estimation will be possible by combining ground and remote sensing information in statistical regression estimation models.
3. The estimates/analyses can be and should be as defensible as possible. A fundamental principle in FIA is to keep things simple: KIS (keep it simple). Our recommendation is to keep the design simple but allow for more complexity in the analyses, since different people want to use the data in different ways. We are likely to have much controversy in analyzing the annualized data sets before agreement is reached.

XII. References

Aldrich, R. C. 1979. Remote sensing of wildland resources: a state-of-the-art review. Gen. Tech. Rep. RM-71. Fort Collins, CO: Rocky Mountain Forest and Range Experiment Station. 56 p.

Arvanitis, L. G.; Reich, R. M. 2004. Natural resources sampling. New York: Sage.

Avery, T. E.; Burkhart, H. E. 1983. Forest measurements. 3d ed. New York: McGraw-Hill. 331 p.

Biggs, P. H.; Pearce, C. J.; Wescott, T. J. 1989. GPS navigation for large-scale photography. Photogrammetric Engineering and Remote Sensing. 55: 1737–1741.

Biggs, P. H.; Spencer, R. D. 1990. New approaches to extensive forest inventory in Western Australia using large-scale aerial photography. Australian Forestry. 53: 182–193.

Bitterlich, W. 1947. The angle count method (in German). Allgemeines Forst-und Holzwirtschaftliche Zeitung. 58: 94–96.

Bolstad, P. V.; Smith, J. L. 1992. Errors in GIS. Journal of Forestry. November: 21–29.

Box, George; Draper, Norman. 1987. Empirical model building and response surfaces. John Wiley & Sons: 74.

Brewer, K. R. W.; Hanif, M. 1983. Sampling with unequal probabilities (lecture notes in statistics). New York: Springer-Verlag. 164 p.

Buckland, S. T.; Anderson, D. R.; Burnham, K. P.; Laake, J. L.; Borchers, D. L.; Thomas, L. 2001. Introduction to distance sampling. Oxford University Press. 432 p.

Bunge, J.; Fitzpatrick, M. 1993. Estimating the number of species: a review. Journal of the American Statistical Association. 88: 364–373.

Burnham, K. P. 1980. Is finite population sampling always applicable to finite populations? Invited presentation to American Statistical Association national meeting; 1980 August; Houston, TX.

Carroll, R. J.; Rupert, D. 1988. Transformations and weighting in regression. New York: Chapman and Hall. 249 p.

Cassel, C-M.; Sarndal, C-E.; Wretman, J. H. 1977. Foundations of inference in survey sampling. New York: John Wiley & Sons. 192 p.

Chao, A.; Lee, S-M. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association. 87: 210–217.

Cochran, W. G. 1977. Sampling techniques. 3d ed. New York: John Wiley & Sons. 428 p.

Congalton, R. G.; Green, K. 1992. The ABCs of GIS. Geographic Information Systems. Part 1. Journal of Forestry. 90(11): 13–20.

Cramer, H. 1963. Mathematical methods of statistics. Princeton, NJ: Princeton University Press. 575 p.

Czaplewski, R. C. 1999. Multistage remote sensing. Towards an annual national inventory. Journal of Forestry. 97(12): 44–48.

Czaplewski, R. C. 2003. Can a sample of Landsat sensor scenes reliably estimate the global extent of tropical deforestation? International Journal of Remote Sensing. 24: 1409–1412.

Dawid, A. P. 1983. Inference, statistical: I. In: Kotz, S.; Johnson, N. L. New York: John Wiley & Sons. Encyclopedia of statistical science. 4: 89–105.

De Vries, P. G. 1986. Sampling theory for forest inventory. A teach-yourself course. New York: Springer-Verlag. 399 p.

Deming, W. E. 1975. On probability as a basis for action. American Statistician. 29: 146–152.

Ducey, M. J.; Gove, J. H.; Valentine, H. T. 2004. A walkthrough solution to the boundary overlap problem. Forest Science (in process).

Duncan, G. J.; Kalton, G.1987. Issues of design and analysis of surveys across time. International Statistical Revue. 55: 97–117.

European Commission. Study on European forestry information and communication system—reports on forestry inventory and survey systems. Vol. 1 and 2. Luxembourg: Office for Official Publications of the European Communities: L-2985.

Feinstein, A. R. 1988. Scientific standards in epidemiological studies of the menace of daily life. Science. 242: 1257–1263.

Franco-Lopez, H. 1999. Updating forest monitoring systems estimates. EM-7140.28. Minneapolis: The University of Minnesota. 48 p. Dissertation.

Fraser, D. A. S. 1983. Inference, statistical: II. In Kotz, S.; Johnson, N. L. New York: John Wiley & Sons. Encyclopedia of statistical science. 4: 105–114.

Freese, F. 1962. Elementary forest sampling. Agriculture Handbook. 232. Washington, DC: U.S. Department of Agriculture, Forest Service. 91 p.

Gadbury, G. L.; Schreuder, H. T. 2003. Cause-effect relationships in analytical surveys: an illustration of statistical issues. Environmental Monitoring and Assessment. 83: 205–227.

Goebel, J. J.; Schreuder, H. T.; House, C. C.; Geisler, P. H.; Olsen, A. R.; Williams, W. W. 1998. Integrating surveys of terrestrial natural resources: the Oregon demonstration project. Tech. Rep. 2. Fort Colins, CO: U.S. Department of Agriculture, Forest Service, Forest Inventory and Monitoring Institute. 20 p.

Green, K. 1992. Spatial imagery and GIS. Journal of Forestry. November: 32–45.

Gregoire, T. G. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Canadian Journal of Forest Research. 28: 1429–1447.

Gregoire, T. G.; Scott, C. T. 1990. Sampling at the stand boundary: a comparison of the statistical performance among eight methods. In: Proceedings XIX World Forestry Congress IUFRO; 1990 August 5–11; Montreal, Canada. Publ. FWS-3-90. Blacksburg: Virginia Polytech Institute and University: 78–85.

Gregoire, T. G.; Valentine, H. T. 2004. Sampling techniques for natural resources and the environment. New York: Chapman Hall CRC Press. In process.

Grosenbaugh, L. R. 1964. Some suggestions for better sample-tree measurement. In: Proceedings Society of American Foresters; 1963; Boston, MA: 36–42.

Grosenbaugh, L. R. 1967. The gains from sample-tree selection with unequal probabilities. Journal of Forestry. 65: 203–206.

Haas, P. J.; Stokes, L. 1998. Estimating the number of classes in a finite population. Journal of the American Statistical Association. 93: 1475–1487.

Hahn, G. J.; Meeker, W. O. 1993. Assumptions for statistical inference. American Statistician. 47: 1–11.

Hajek, J. 1957. Some contributions to the theory of probability sampling. Bulletin of the International Statistical Institute. 36: 127–133.

Hansen, M. H.; Hurwitz, W. N.; Madow, W. G. 1953. Sample survey methods and theory. Vol. I and II. New York: John Wiley & Sons. 638 p, 332 p.

Hill, A. B. 1965. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine. 58: 295–300.

Holmgren, P.; Thuresson, T. 1998. Satellite remote sensing for forestry planning: a review. Scandinavian Journal of Forest Research. 13: 90–110.

Hush, B. 1971. Planning a forest inventory. FAO Forest Products Studies No. 17. Rome, Italy. 121 p.

Iles, K. 2003. A sampler of inventory topics. Nanaimo, B.C., Canada: Kim Iles Associates. 869 p.

Johnson, E. W. 2000. Forest sampling desk reference. New York: CRC Press. 985 p.

Kalton, G.; Anderson, D. W. 1986. Sampling rare populations. Journal of the Royal Statistical Society A. 149: 65–82.

Kish, L. 1967. Survey sampling. 2d ed. New York: John Wiley & Sons. 643 p.

Koch, G. G.; Gillings, D. B. 1983. Inference, design based vs model based. In: Kotz, S.; Johnson, N. L. New York: John Wiley & Sons. Encyclopedia of statistical science. 4: 84–88.

Kotz, S.; Johnson, N. L. New York: John Wiley & Sons. Encyclopedia of statistical science 8. 870 p.

Kruskal, W. H.; Mosteller, F. 1979. Representative sampling. In: Kotz, S.; Johnson, N. L. New York: John Wiley & Sons. Encyclopedia of statistical science. 8: 77–88.

Kutner, M.; Neter, J.; Nachtsheim, C.; Wasserman, W. 2003. Applied linear regression models. 4th ed. New York: McGraw-Hill/Irwin. 672 p.

Lachowski, H.; Maus, P.; Platt, B. 1992. Integrating remote sensing with GIS. Journal of Forestry. 12: 16–21.

Lefsky, M. A.; Cohen, W. B.; Parker, G. G.; Harding, D. J. 2002. Lidar remote sensing for ecosystem studies. Bioscience. 52: 19–30.

Lillesand, T. M.; Kiefer, R. W. 1987. Remote sensing and image interpretation. 2d ed. New York: John Wiley & Sons. 721 p.

Lin, J-M. 2003. Small area estimation. Fort Collins: Colorado State University, Statistics Department. 344 p. Dissertation.

Max, T. A.; Schreuder, H. T.; Hazard, J. W.; Oswald, D. D.; Teply, J.; Alegria, J. 1996. The Pacific Northwest Region Vegetation and Inventory Monitoring System. Res. Pap. PNW-RP-493. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Mosteller, F.; Tukey, J. W. 1977. Data analysis and regression. Reading, MA: Addison-Wesley Publishing Co. 586 p.

Murthy, M. N. 1967. Sampling theory and methods. Calcutta, India: Statistical Publishing Co. 684 p.

Olsen, A. R.; Schreuder, H. T. 1997. Perspectives on large-scale natural resource surveys when cause-effect is a potential issue. Environmental and Ecological Statistics. 4: 167–180.

Overton, W. S.; Stehman, S. V. 1995. The Horvitz-Thompson Theorem as a unifying perspective for probability sampling: with examples from natural resource sampling. American Statistician. 49: 261–268.

Peifer, M. 1997. Cancer-beta-catenin as oncogene: the smoking gun. Science. 75: 1752–1753.

Pinkham, R. S. 1987. An efficient algorithm for drawing a simple random sample. Applied Statistics. 36: 370–372.

Rosenfield, G. H.; Fitzpatrick-Lins, K. 1986. A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing. 52: 223–227.

Sarndal, C-E. 1980. A two-way classification of regression estimation strategies in probability sampling. Canadian Journal of Statistics. 8: 165–177.

Sarndal, C-E.; Swensson, B.; Wretman, J. 1992. Model assisted survey sampling. New York: Springer-Verlag. 694 p.

Schreuder, H. T. 1994. Simplicity versus efficiency in sampling designs and estimation. Environmental Monitoring and Assessment. 33: 237–245.

Schreuder, H. T.; Alegria, J. 1995. Stratification and plot selection rules, misuses and consequences. Res. Note RM-RN-536. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station. 4 p.

Schreuder, H. T.; Bain, S.; Czaplewski, R. C. 2003. Accuracy assessment of percent canopy cover, cover type and size class. Gen. Tech. Rep. RMRS-GTR-108. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 10 p.

Schreuder, H. T.; Czaplewski, R. L. 1992. Long-term strategy for the statistical design of a forest health monitoring system. Environmental Monitoring and Assessment. 27: 81–94.

Schreuder, H. T.; Geissler, P. H. 1999. Plot designs for ecological monitoring of forest and range. North American Science Symposium. Towards a unified framework for inventorying and monitoring forest ecosystem resources symposium; 1998 November 1–4; Guadalajara, Mexico. Proc. RMRS-P-2. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station: 180–185.

Schreuder, H. T.; Gregoire, T. G. 2001. For what applications can probability and non-probability sampling be used? Environmental Monitoring and Assessment. 66: 281–291.

Schreuder, H. T.; Gregoire, T. G.; Wood, G. B. 1993. Sampling methods for multiresource forest inventory. New York: John Wiley & Sons. 446 p.

Schreuder, H. T.; Li, H. G.; Sadooghi-Alvandi, S. M. 1990. Sunter's pps without replacement sampling as an alternative to Poisson sampling. Res. Pap. RMRS-RP-290. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station. 6 p.

Schreuder, H. T.; Lin, J-M. S.; Teply, J. 2000. Estimating the number of tree species in forest populations using current vegetation survey and forest inventory and analysis approximation plots and grid intensities. Res. Note RMRS-RN-8. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 7 p.

Schreuder, H. T.; Schreiner, D. A.; Max, T. E. 1981. Ensuring an adequate sample at each location in point sampling. Forest Science. 27: 567–578.

Schreuder, H. T.; Thomas, C. E. 1991. Establishing cause-effect relationships using forest survey data. Forest Science. 37: 1497–1525.

Schreuder, H. T.; Williams, M. S. 2000. Reliability of confidence intervals calculated by bootstrap and classical methods using the FIA 1-ha plot design. Gen. Tech. Rep. RMRS-GTR-57. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 6 p.

Schreuder, H. T.; Williams, M. S.; Reich, R. 1999. Estimating the number of tree species in a forest community using survey data. Environmental Monitoring and Assessment. 56: 293–303.

Schwarz, C. J.; Seber, G. A. F.1999. Estimating animal abundance. Review III. Statistical Science. 14: 427–456.

Shiver, B. D.; Borders, B. E. 1996. Sampling techniques for forest resource inventory. New York: John Wiley & Sons. 356 p.

Smith, T. M. F. 1994. Sample surveys: 1975–1990; an age of reconciliation? International Statistical Review. 62: 5–34.

Stahl, G.; Ringvall, A.; Fridman, J. 2001. Assessment of coarse woody debris—a methodological overview. Ecological Bulletin. 49: 57–70.

Stuart, A. 1964. Some remarks on sampling with unequal probabilities. Bulletin of the International Statistical Institute. 40: 773–780.

Taubes, G. 1995. Epidemiology faces its limits. Special news report. Science. 269: 164–169.

Thompson, W. L.; White, G. C.; Gowan, C. 1998. Monitoring vertebrate populations. New York: Academic Press. 365 p.

USDA Forest Service. 1998. Implementation of remote sensing for ecosystem management. Salt Lake City, Utah. U.S. Department of agriculture, Forest Service, Engineering Staff, Remote Sensing Applications Center.

Wensel, L.; Levitan, J.; Barber, K. 1980. Selection of basal area factor in point sampling. Journal of Forestry. 78: 83–84.

Williams, M. S.; Gove, J. H. 2003. Perpendicular distance sampling: an alternative method for sampling downed coarse woody debris. Canadian Journal of Forest Research. 33: 1–16.

Wood, G. B. 1988. Generating the list of random numbers for 3P samples. Australian Forester. 50: 260–264.

Wood, G. B. 1990. Ground sampling methods used to inventory tropical mixed/moist forest. Forest Ecology and Management. 35: 199–206.

XIII. Glossary

Accuracy. Freedom from error or the closeness of a measurement or estimate to the true value. More broadly, it is the degree to which a statement or quantitative result approaches the truth. Note that $\text{Accuracy} = \sqrt{\text{Precision}^2 + \text{Bias}^2}$ using these statistical definitions. Thus, if bias is eliminated, Accuracy = Precision.

Asymptotically unbiased. Estimation bias goes to 0 as sample size approaches population size. It is the same as consistency as used by Cochran (1977).

Attribute. Units classified as having or not having some specific quality.

Basal area (per site). The cross-sectional area at breast height of all trees on the site.

Basal area (per tree). The cross-sectional area of a tree at breast height.

Bias. A systematic error introduced into sampling, measurement, or estimation by selecting or favoring, possibly unintentionally, one outcome or answer over others.

Breast height. The point on a tree stem at 1.4 m (4'6") in the USA, New Zealand, Burma, India, Malaysia, South Africa, and some other countries and 1.3 m (4'3") above ground in continental Europe, Great Britain, Australia, Canada, and Mexico.

Consistency. The same as asymptotically unbiased as defined above.

Continuous variable. A variable expressed in a numerical scale of measurement, where any interval of it can be subdivided into an infinite number of values.

Correlation coefficient. A measure of the degree of linear association between two variables that is unaffected by the sizes or scales of the variables.

Covariance. A variance or measure of association between paired measurements of two variables.

Covariate. A quantitative, often explanatory variable in a model such as a regression model. Covariates are often important in improving estimation.

dbh. The diameter at breast height of a tree.

Discrete variable. Qualitative variables or those represented by integral values or ratios of integral values.

Double sampling. Two levels of sampling where the first level provides information on covariates and the second on the variable of interest to estimate parameter(s).

Efficient estimator. An estimator that predicts a parameter more reliably than competing estimators where reliability is usually measured by the ratio of the mean square errors of the estimators.

Estimate. The numerical value calculated from an estimator for a sample.

Estimator. A function of the values in a sample or a formula used for estimating a parameter based on a sample.

Estimator of population mean. The formula used in estimating the population mean from a sample.

Estimator of population variance. The formula used in estimating the population variance from a sample.

Experiment. The conduct of a systematic, controlled test or investigation.

Global positioning system (GPS). A system using satellites to locate ground positions.

Inference. The drawing of conclusions based on data or observations.

Mean. The average value of a variable for all units in a population or sample.

Median. The value of a variable so that half of the values are larger and half are smaller than this value in a population or sample.

Mode. The value of a variable that occurs most frequently in a population or sample.

Multilevel sampling. A sampling design where more than one phase or stage of sampling is used. The first levels are used to collect information on covariates useful for more efficient estimation of the ultimate parameter(s) of interest for which information is usually collected at the last phase or stage.

Parameter. A characteristic or function of the values of the units in a population, i.e., the population characteristic of interest, such as average volume per ha or total volume of trees in a forest.

Population. An aggregate of items each with a common characteristic or common set of characteristics. In the statistical sense, a population is an assembly of individual units formed in order to describe the population quantitatively. For example, it might be all the trees in a particular forest stand or all the users of a recreation area.

pps sampling. A sampling design where sample units are selected with a probability proportional to a measure of size, usually a covariate such as dbh or basal area in the case of tree volume.

Precision. Relative freedom from random variation. In sampling it is expressed as the standard error of the estimate and relates to the degree of clustering of sample values about their own average or the reproducibility of an estimate in repeated sampling. It is also used to indicate the resolving power of a measuring device.

Probabilistic sampling. Procedures in which samples are selected such that all units and each pair of units in the population have a positive probability of selection.

Randomization. A deliberately haphazard arrangement of observations to simulate selection by chance.

Sample surveys. The design and execution of surveys to provide estimates of characteristics (parameters) of well-defined finite populations.

Sample unit. A unit from a population, i.e., a tree or all trees located within a plot (i.e., fixed-area, strip or point sample).

Sample. A subset of a population used to obtain estimates of one or more of its parameters. In this book we focus on probabilistic samples. For example, a sample can be the diameters (dbh) of all trees on a sample of plots or the amount of time spent picnicking by users of a recreation area on given days.

Sampling design. A formalized method of selecting a sample from the population, for example simple random sampling.

Sampling frame. A list of all sample units used to represent a population.

Sampling strategy. Comprises both the sampling design and the estimator(s) used, for example simple random sampling with the estimator of the population mean, say the sample mean.

Single-level sampling. A sampling design where units are selected directly from the sampling frame of the population.

Standard deviation. The square root of the variance defined below.

Statistical inference. Expressing the connection between the unknown “state of nature” and observed information in probabilistic terms.

Statistical survey. Involves the design and execution of surveys to provide estimates of characteristics of well-defined finite populations.

Unequal probability sampling. Sampling designs where units are selected with different probabilities. These probabilities need to be known for unbiased estimation.

Unit. The basic sample unit used; e.g., that used in the last stage of multistage sampling.

Variable. A characteristic that varies from unit to unit; for example, the age of a tree.

Variance. The average of the deviations squared between the values of the variables and the overall mean in the case of a population or between the values of the variables and the sample mean in the case of a sample; in the first case it is a population parameter, in the second a sample statistic.

Appendix 1. Inference

Inductive logic, the drawing of conclusions from analysis of observed data about unobserved parameters or underlying laws, is one of the most controversial issues in philosophy (Gregoire 1998, Schreuder and Gregoire 2001). Although inference, the drawing of conclusions based on data or observations, is not limited to the narrow field of scientific and statistical inference, the latter is important in this contentious world, and a proper understanding of it is crucial to discussing the role of sampling in the inferential process. Scientific inference becomes statistical inference when the connection between the unknown "state of nature" and the observed information is expressed in probabilistic terms (Dawid 1983).

Statistical inference comprises the whole field of statistics, its focus being what is logically implied by the information available (Fraser 1983). Cramer (1946) summarizes the role of statistical inference as having three functions: description, analysis, and prediction. Description is the reduction of data sets to as small a set of numbers as possible, such as the mean, variance, skewness of a distribution, etc. This enables one to describe a population as concisely and briefly as possible and can allow for comparison between populations. Analysis is the summarization of data for a particular purpose or objective. Examples are: What are the estimates of certain population characteristics? Did the sample arise from a given distribution? Or given two samples, did they arise from the same population or not? Statistics provides methods of how to do such analyses. Statistical methods are used to predict and explain phenomena, often a very challenging task.

Ideally, statistical inference would always be based on Bayes theorem, which combines prior information with information from surveys or experiments and would be acceptable to many statisticians if the prior belief is objective. The problem is that usually prior information is subjective, where subjective indicates that the information available varies from person to person. Objective prior information indicates that people would normally agree on it. As an example of subjective prior information, a forest industry person could believe that there is plenty of old growth distributed nicely over the forest for habitat for endangered species whereas an environmentalist could equally strongly believe that the old growth in the forest is limited and badly distributed. People willing to accept prior subjective information are called Bayesians and rely on Bayes theorem for inference. Non-Bayesians or frequentists, a majority, use classical inference procedures relying only on objective data often based on normality assumptions and large sample theory based on the central limit theorem and related statistical properties. It is our belief that Bayesian procedures should be used when immediate logically defensible decisions need to be made, and classical ones when building a body of scientific knowledge. A forest manager who has to make decisions about whether or not to cut old growth and where for management purposes, may well choose to use all his prior information to construct a (subjective) prior distribution to combine with actual sample data in order to use Bayes theorem to make such decisions. Such decisions can be defended at least on the basis of a systematic approach. Scientific databases can be used by different users applying different priors to make their decisions.

Statistical inference from sample surveys can be either model-based or design-based. In model-based sampling, inference relies on a statistical model to describe how the probability structure of the observed data depends on uncontrollable chance variables and, frequently, on other unknown nuisance variables. Such models can be based on a theoretical understanding of the process by which the data were generated, experimental techniques used, or past experience with similar processes. For inference in design-based sampling, reliance is placed on probabilistic sampling. It is the most widely accepted approach now. The following is a brief summary of both approaches.

In non-probabilistic or model-based sampling, inference is made by specifying an underlying superpopulation model ξ for the values of the variable in the actual population being sampled. The actual values are considered to be random variables from this superpopulation. It is then assumed that the actual population or a sample from it is a sample from this superpopulation of interest. Then using ξ, for estimator $\hat{Y}$ of the quantity of interest Y, the distribution of $\hat{Y}$-Y can be derived for the specific sample and the model-based mean square error of $\hat{Y}$-Y can be obtained and estimated, leading to a model-based prediction interval for Y. Inference extends to parameters of the superpopulation model, so that the inference space is broader than for design-based inference. Sample units do not have to be chosen at random or with known probability as long as they are not selected based on their values of interest $y_i,\ i = 1,...,N$.

Conclusions and inferences rely heavily on the model assumed, which can be a serious liability if the model is not specified correctly. But if a model is correctly specified, an increase in precision can be expected over the design-based approach. Our experience is that very few models are reliable. The statement by Box and Draper (1987) that "all models are wrong; the practical question is how wrong do they have to be to not be useful" deserves consideration. Nonetheless, models are useful in building a body of knowledge in every subject and may have to be relied on when a quick decision needs to be made. The Bayesian approach to inference with subjective priors fits well into the model-based inference approach, although many advocates of model-based sampling would not consider themselves Bayesians at all. As noted by Koch and Gillings (1983), model-based inference encompasses Bayesian and superpopulation inference since the validity of the claimed generality is model dependent, i.e., is sensitive to model misspecification.

The design-based approach to inference relies heavily on probabilistic sampling, in which each sample unit of the population has a positive probability of being selected and the probability of each sample can be calculated. The statistical behavior of estimators of a population attribute is based on these probabilities and the probability-weighted distribution of all possible sample estimates. The distribution of the variable, probabilistic or otherwise, is not considered here. An obvious weakness of this approach is that samples that were not drawn are considered heavily in evaluating the properties of the inference procedure, yet should not inference about a population parameter be based solely on the actual sample drawn? Nevertheless, the approach is objective and the only assumption made is that observed units are selected at random so the validity of the inference only requires that the targeted and sampled populations are the same. And, careful attention to sample selection within the framework of probabilistic sampling will eliminate some undesirable samples from consideration and give others a low probability of selection. The whole idea behind probabilistic sampling is to make the sample representative of the population being sampled. However, as illustrated especially well in Kruskal and Mosteller (1979), "representative" is subject to a wide array of interpretations.

Smith (1994, p.17), formerly a strong advocate of model-based inference, states: "My view is that there is no single right method of inference. All inferences are the product of man's imagination and there can be no absolutely correct method of inductive reasoning. Different types of inference are relevant for different problems and frequently the approach recommended reflects the statistician's background such as science, industry, social sciences or government…I now find the case for hard-line randomization inference based on the unconditional distribution to be acceptable…Complete reconciliation is neither possible nor desirable. Vive la difference."

A crucial difference between design-based and model-based approaches is that for the former inference is made about the finite, usually large population sampled, whereas model-based sampling makes inferences about superpopulations by the compulsory use of models. The inference then is about the actual population that is represented in some sense by the existing population, assuming that the models used underlie the real population.

Deming (1975) recommends that a distinction should be made between enumerative (or descriptive) and analytical (or comparative) surveys. In enumerative surveys interest is in a finite, identifiable, and unchanging population from which samples are drawn. Action is taken on the population of units studied (e.g., all forests in the state of Montana) at the time of sampling to decide how much timber to harvest. This is the type of survey conducted by Forest Inventory and Analysis (FIA) program of the USDA FS. Here design-based inference is indicated. In contrast, analytical surveys focus on populations where action is to be taken on the process or cause system, the purpose being to improve conditions in the future. For the Montana forests, we are still talking about the same forests but different conditions exist when we apply treatments to improve conditions. For example, land management agencies such as the National Forest System (NFS) of the US Forest Service may be interested in collecting information on managing rare and endangered wildlife species so as to create or modify existing vegetation conditions to increase the number of such animals in the forests over time. Although we still want to take a design-based sample here from the existing population, the inference clearly is for populations of the future so is model-based in the sense that we are extrapolating from the existing to future populations. This is clarified further by the following.

Deming (1975) suggests that in enumerative surveys a 100 percent sample of the population provides the complete answer to the questions posed, whereas in analytical surveys the answer is still inconclusive. Hahn and Meeker (1993) make the further distinction that analytical studies require the assumption, usually not verifiable, that the process about which one wants to make inferences is statistically identical to that from which the sample was selected. Figure A-1 from Hahn and Meeker (1993) illustrates the differences between analytical and enumerative surveys. A problem with this distinction is that often in surveys conducted by land management agencies, there is interest in both types of inference. For example: one aim may be to determine how much timber to harvest and from where or what areas to delineate for weed or erosion control (requiring enumerative surveys) and another to assess what needs to be done to improve habitat for wildlife or diminish it for noxious weeds (requiring analytical surveys).

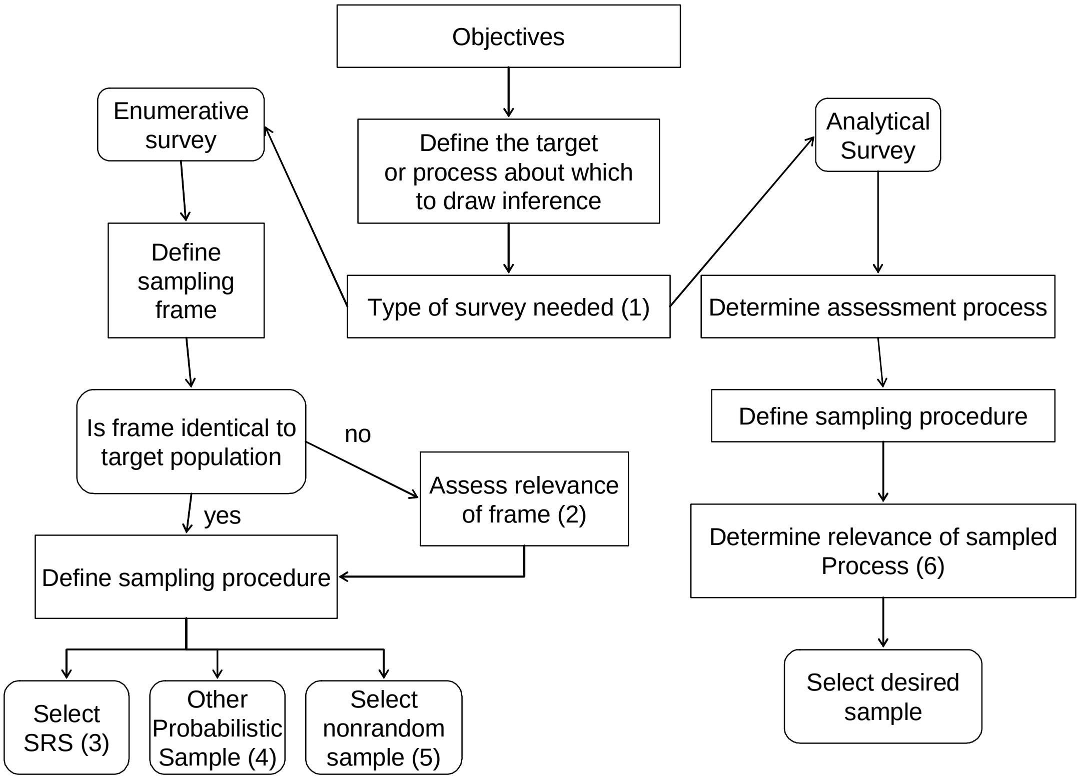

Objectives
Enumerative survey
Define the target or process about which to draw inference
Analytical Survey
Define sampling frame
Type of survey needed (1)
Determine assessment process
Is frame identical to target population
no
Define sampling procedure
Assess relevance of frame (2)
yes
Determine relevance of sampled Process (6)
Define sampling procedure
Select desired sample
Select SRS (3)
Other Probabilistic Sample (4)
Select nonrandom sample (5)

Figure A-1. A Comparison of Enumerative and Analytical Surveys. Reprinted with permission from *The American Statistician.* Copyright 1993 by the American Statistical Association. All rights reserved. The numbers refer to the following comments:

(1) Is the purpose of the study to draw conclusions about an existing finite population (enumerative study), or is it to act on and/or predict the performance of a (frequently future) process (analytic study)?

(2) Statistical intervals apply to the frame from which the sample is taken. When the frame does not correspond to the target population, inferences about the target population could be biased, and a statistical interval provides only a lower bound on the total uncertainty.

(3) Most statistical intervals assume a simple random sample from the frame.

(4) More complex statistical intervals than those for simple random samples apply; see Cochran (1977).

(5) Statistical intervals do not apply. If they are calculated, they generally provide only a lower bound on the total uncertainty.

(6) Statistical intervals apply to the sampled process, and not necessarily to the process of interest. Thus, any statistical interval generally provides only a lower bound on the total uncertainty with regard to the process of interest.

Appendix 2. Distributions

Populations are discrete or finite ($N < \infty$) but often are assumed to be infinitely large $(N = \infty)$ since continuous distributions are more likely to approximate the real population. Infinite populations have properties that are crucial to statistical inference. A few of the more important distributions are presented below as well as some key results from statistical theory based on continuous distributions. The material is a condensation of material discussed in Schreuder and others (1993).

Distributions are often characterized by their moment generating function (mgf).

Definition. If *Y* is a random variable with probability density f(y), then the expected value, *E*, of e^{ty} is called the mgf of *Y* if it exists for every value of *t* in some interval $-h^2 < t < h^2$. This is denoted by

$$m(t) = E(e^{ty}) = \sum_{y} e^{ty} f(y)dy$$

where for discrete distributions, *f*(*y*) is the probability mass function, and

$$m(t) = E(e^{ty}) = \int_{-\infty}^{\infty} e^{ty} f(y)dy \text{ where}$$

for continuous distributions, *f*(*y*) is the probability density function (Mood and others 1974). The logarithm of the mgf, called the cumulant generating function, is often used. The moments of this function are called the cumulants.

m(*t*) generates the moments of distributions. For example, in survey sampling we are often interested in estimating the first two moments of the normal distribution. The first moment is the mean and is obtained from the mgf by differentiating with respect to *t once* and setting *t* = 0. Similarly the second moment, the variance, is obtained by differentiating the mgf with respect to *t* twice and setting *t* = 0 and then subtracting the first moment squared at *t* = 0.

Continuous Distributions

Three important continuous distributions are the normal, gamma, and multivariate.

Normal Distribution

The cumulative distribution, usually simply called distribution, is defined as

$$F(y) = P\{Y \le y\} = \int_{-\infty}^{y} (1/\sqrt{2\pi\sigma} \exp[-1/2\{(y-\mu)/\sigma\}^2]dy = \int_{-\infty}^{y} f(y;\mu,\sigma)dy$$

with parameters μ and σ^2 $(\sigma > 0)$, and $-\infty < Y < \infty$. The parameters μ and σ^2 are called the mean and variance of the distribution, respectively.

The mgf of the normal distribution is

$$m(t) = \exp(t\mu + t^2\sigma^2/2)$$

with mean $m'(0) = \mu_1' = \mu$ and variance $m''(0) - [m'(0)]^2 = \mu_2 = \sigma^2$ so in fact the mean and variance are the two parameters of the distribution. Although there are numerous situations where the normal distribution approximates the distribution of units in a population (such as the heights of all trees in a large plantation), it is most commonly used as a convenient approximation to other distributions. The normal distribution is important in probability theory because it is the limiting distribution of almost any standardized sums (or means) of random variables as the number of variables in the sum increases.

If a statistic $\hat{\theta}$ is an unbiased estimator of a parameter θ, with estimated variance $v(\hat{\theta})$, and is approximately normally distributed, then the statistic $t = (\hat{\theta} - \theta)/\sqrt{v(\hat{\theta})}$ follows Student's t-distribution with density $f(t) = \tau\{(v+1)/2)\}(1/\sqrt{v})(1+t^2/v)^{-(v+1)/2}/\{\tau(v/2)\tau(1/2)\}$ where *v* is the number of degrees of freedom on which the estimate of the standard error is based and $\tau(v/2) = \int_0^\infty t^{v/2-1}e^{-t}dt$. The t-distribution is fundamental to constructing confidence intervals, and tables for this distribution are widely available (see Appendix 3, Table 2).

Gamma Distribution

This distribution appears naturally as the distribution of the sum of squares of independent, standard, normally distributed random variables, $Z_1, Z_2, ..., Z_n$. Then $\sum_{i=1}^{n} Z_i^2$ has a χ^2 distribution with parameter n where χ^2 is a special case of the gamma distribution and n is the number of degrees of freedom. The gamma distribution is often used in survey sampling to enable comparisons of sampling strategies. The distribution is:

$$F(y) = P[Y \le y] = \int_0^y y^{\alpha-1}\exp(-y/\beta)/\{\beta^\alpha \tau(\alpha-1)\}dy = \int_0^y f(y)dy$$

with parameters α and $\beta > 0$, y>0, and $\tau(\alpha-1) = \int_0^y t^{(\alpha-2)}e^{-t}dt$.

The gamma distribution has the mgf

$$m(t) = (1-\beta t)^{-\alpha}$$

with mean $\mu = \alpha\beta$ and variance $\mu_2 = \alpha\beta^2$.

Multivariate Distributions

The multivariate normal distribution has been studied much more extensively than other multivariate distributions and is used more frequently for inference among multivariate continuous distributions than is the normal distribution among univariate continuous distributions. The bivariate normal distribution is defined as

$$F(x,y) = P(X \le x, Y \le y) = (2\pi\sigma_x\sigma_y\sqrt{1-\rho^2})^{-1}\exp\{[-1/\{2(1-\rho^2)\}]\{(x-\mu_x)^2/\sigma_x^2 - 2\rho\{(x-\mu_x)/\sigma_x\}$$
$$\{(y-\mu_y)/\sigma_y + (y-\mu_y)^2/\sigma_y^2\}]$$

with $-\infty < x < \infty, -\infty < y < \infty$, where $\mu_x = E(X), u_y = E(Y), \sigma_x^2 = V(X), \sigma_y^2 = V(Y)$,

and $\rho = \sqrt{E(X-\mu_x})(Y-\mu_y)/(\sigma_x^2\sigma_y^2) = \sigma_{xy}/(\sigma_x\sigma_y), (-1 < \rho < 1)$ is called the correlation between *X* and *Y*,

and σ_{xy} is the covariance between *X* and *Y*.

Discrete Distributions

Important examples of discrete distributions are the binomial, hypergeometric, Poisson, and multinomial distributions.

Binomial Distribution

If n independent trials are made (such as whether a load of logs should be sampled or not) and each trial has the probability p of outcome i occurring, then the number of times in which i occurs may be represented as a random variable *Y* following the binomial distribution with parameters n and *p*. This distribution is defined as the distribution of a random variable *Y* (= number of occurrences of *i*) for which

$$P[Y = y] = [n!/\{y!(n-y)!\}]p^y(1-p)^{n-y} \quad (y = 0,1,2,\ldots,n).$$

The mgf of the binomial distribution is

$$m(t) = (1 - p + pe^t)^n \, .$$

From the mgf, the mean and variance are derived as

$$\mu = np, \mu_2 = \mu_2^{'} - (\mu_1^{'})^2 = p(1-p) \text{ where } \mu_2^{'} = np + n(n-1)p^2 \, .$$

There are many approximations to the binomial distribution. Such approximations often involve limiting distributions that arise when one or both parameters converge to a specific value. Limiting distributions are the (discrete) Poisson distribution discussed below $(n \to \infty, p \to 0)$ with *np = a* constant θ, and the normal distribution, which is a special case of the standardized binomial variable $(Y - np)/\sqrt{np(1-p)}$ as $n \to \infty$.

Hypergeometric Distribution

The classic situation in which it arises in forestry is as follows.

Suppose we have a population with *N* trees, *M* of which are dead, and *N* - *M* of which are alive. If n trees are drawn at random without replacement, then the probability of selecting *y* dead trees is

$$P[Y = y] = \{M!/[y!(M-y)!]\}\{(N-M)!/[(N-M-n+y)!(n-y)!]\}/[N!/\{n!(N-n)!\}]$$

for $\max(0, n-N+M) \le y \le \min(n, M)$ with parameters *M*, *N*, and n and n! = n(n-1)(n-2)…1 and 0! = 1.

The mgf for the hypergeometric distribution is

$$m(t) = [(N-n)!(N-M)!/N!]H(-n, -M; N-M-n+1; e^t)$$

where $H(\alpha, \beta; \gamma; z) = 1 + (\alpha\beta/\gamma)(z/1!) + [\alpha(\alpha+1)\beta(\beta+1)]/\{\gamma(\gamma+1)\}z^2/2! + \ldots$

is a hypergeometric function which converges for absolute value of z<1.

The mean is $\mu = nM/N$ and variance $\mu_2 = [(N-n)/(N-1)]n(M/N)(1-M/N)$.

Several approximations to the hypergeometric distribution exist, a simple one being the binomial distribution $P[Y = y] = [n!/\{y!(n-y)!\}]M/N)^y(1-M/N)^{n-y}$ which is usually adequate when *n* < 0.1*N*. Confidence intervals can be constructed as described under the binomial distribution assuming either the binomial or normal approximation to the hypergeometric distribution.

Poisson Distribution

If the future lifetime of an item of equipment (say a chainsaw) is independent of its present age, then the lifetime can be represented by a random variable *Y* with distribution $P[Y = y] = e^{-\theta}\theta^y/y!, y = 0,1,2,\ldots; \theta > 0$ where θ is the only parameter (= *np* in the binomial distribution with $n \to \infty$ as $p \to 0$). A widely quoted application of this distribution concerns the number of soldiers annually kicked to death by mules in an army at the middle of the 19th century. (An analogous situation might be the number of loggers killed by falling trees in a forest.) The probability of death was small and the number of soldiers exposed was large. It is doubtful that the conditions of independence and constant probability (*p*) were satisfied but the data available were satisfactorily fitted by this distribution.

The moment generating distribution is $m(t) = \exp[\theta(e^t - 1)]$

with mean $\mu = \theta$ and variance $\mu_2 = \theta$.

Multinomial Distribution

Multivariate discrete distributions are often closely related to univariate ones. For example, the marginal distribution of individual variables is generally a simple binomial, Poisson, or hypergeometric distribution or a distribution obtained by modifying or compounding one of the univariate distributions. For example, the joint distribution of the random variables $n_1, n_2, ..., n_k$ representing the number of occurrences of events $O_1, O_2, ..., O_k$ in n trials is the multinomial distribution

$$P(n_1, n_2, ..., n_k) = n!\Pi_{j=1}^{k}(p_j^{n_j}/n_j!)$$

with $0 \le n_j$ for all j = 1, . . . , k and $\sum_{j=1}^{k} n_j = n$. This distribution is a natural extension of the binomial distribution, which is a special case if k = 2. The joint distribution of any subset s < k is also a multinomial, hence the important subset where we have two classes, that is, "class 1" and "all others," is also a binomial.

The mgf of the multinomial is $m(t_1, ..., t_{k)} = (p_1e^{t_1} + ... + p_ke^{t_k})^n$ where the moments of the $n_i (i = 1, ..., k)$ are simply

$$\mu_1^{'}(t) = p_i, i = 1, ..., k$$

and

$$\mu_2(t_i) = p_i(1 - p_i).$$

Note that the covariance of n_i and n_j and the correlation between them are respectively

$$Cov(n_i, n_j) = -np_ip_j$$

and

$$Cor(n_i, n_j) = -\sqrt{p_i}p_j/\{(1 - p_i)(1 - p_j)$$

Confidence intervals for any p_i or n_ip_i are constructed by treating the multinomial as a binomial with probability of selecting class *i*, p_i, and of all other classes, $1 - p_i$. Then the discussion for constructing confidence intervals under the binomial is appropriate.

Multivariate Hypergeometric Distribution

A generalization of the hypergeometric distribution is the multivariate hypergeometric distribution defined as follows: If there is a population of *N* units, N_i of which are of type i *(i* = 1, . . . , *k*) so that $\sum_{i=1}^{k} N_i = N$ and a sample of size n is taken without replacement from the N units, then

$$P(n_1, n_2, ..., n_k) = \prod_{i=1}^{k}\{N_i!/[(N_i - n_i)!n_i!]\}/[N!/\{(N - n)!n!\}]$$

is the multivariate hypergeometric distribution with

$$\sum_{i=1}^{k} n_i = n; 0 \le n_i \le N_i, i = 1, ..., k$$

The moments of the multivariate hypergeometric distribution are analogous to those for the hypergeometric distribution and the correlation between n_i and n_j is

$$Corr(n_i, n_j) = -\sqrt{N_i}N_j/\{(N - N_i)(N - N_j)\}.$$

Laws of Large Numbers

In inductive inference we determine something about a population of interest, say its mean, by examining a sample from the population. The following theorems assuming random sampling are critical to survey sampling inference. A finite number of values of *Y* can be used to make reliable inferences about E(Y), the average of an infinite (or very large finite) number of values of *Y* (i.e., $\sum_{i=1}^{N} y_i / N$ the average for the whole population). For simple random samples, the following three theorems apply:

(1) Theorem 1(Tchebysheff's Inequality)

For a distribution *F*(*y*) with mean μ and finite variance σ^2, and if $\bar{y}$ is the mean of a random sample of size n from this distribution and α any positive number, then

$$P[-\alpha\sigma/\sqrt{n} \le \bar{y} - \mu \le \alpha\sigma/\sqrt{n}] \ge 1 - 1/\alpha^2 .$$

(2) Theorem 2 (Weak Law of Large Numbers)

Let *F(y)* be a distribution with mean μ and finite variance σ^2 and let ε and δ be two specified small numbers where $\varepsilon > 0$ and $0 < \delta < 1$. If n is any integer greater than $\sigma^2/(\varepsilon^2\delta)$ and $\bar{y}_n$ is the mean of a random sample of size *n* from *F*(*y*), then

$$P[-\varepsilon < \bar{y}_n - \mu < \varepsilon] \ge 1 - \delta .$$

Thus this theorem states that for any two small numbers ε and δ, where $\varepsilon > 0$ and $0 < \delta < 1$, there is an integer n such that for a random sample of size n or larger from the distribution of the population of *y*-values *F*(*y*), the probability that the mean of the sample of *y*-values $\bar{y}_n$ is arbitrarily close to the population mean μ can be made as close to 1 as desired. The weak law of large numbers is an example of convergence in probability. The following theorem exemplifies the notion of convergence in distribution. It underlies the wide application of the t-distribution in constructing confidence intervals around estimates of parameters of interest and highlights the critical importance of the normal distribution.

(3) Theorem 3 (Central Limit Theorem)

Let *F*(*y*) be a distribution with mean μ and finite variance σ^2 and let $\bar{y}_n$ be the mean of a random sample of size n from *F*(*y*). *If* $z_n = (\bar{y}_n - \mu)\sqrt{n}/\sigma$, then the distribution of z_n approaches the standard normal distribution as n increases without bound.

This theorem states that the mean $\bar{y}_n$ of a random sample from any distribution with finite mean μ and variance σ^2 is approximately distributed as a normal variable with mean μ and finite variance σ^2/n. Since we usually deal with populations of size *N* with *N* considerably less than ∞, *n* cannot increase without bound, so the applicability of this theorem to finite populations is arguable. In many cases, when n is not too small and the distribution of *y* is not too far from symmetry, the distribution of z_n will be approximately normal for inference.

Appendix 3. Tables

Tabulated values are frequently used in data analysis and hypothesis testing. Among the more common tables (included here) are: Table of random numbers, Distribution of Student's t, confidence intervals (95 percent) for the binomial distribution, ArcSine transformation, and two-tailed significance levels of correlation coefficient *r*.

The tables presented here have been generated using the open source package *R*. The code used to create these tables can be downloaded from http://www.r-project.org/ and can be modified and used for more detailed tables to meet specific needs.

Table 1. Table of random numbers.

55862	47692	92962	37452	24651	87211	80143	24086	72731	65043
97923	60028	26117	73239	61498	11036	01350	60940	45478	60901
17929	18397	91162	93398	56598	01268	68729	94782	04323	25374
88428	55580	07083	85487	57816	64538	86549	25001	79018	20327
94726	02973	87423	03166	20079	63784	07889	05779	66550	07533
33594	96394	84905	60462	20100	83140	26129	16337	74811	65944
38484	42163	27173	84056	59407	91873	92328	39655	49040	90167
23393	82742	38862	26980	75368	02847	01053	75732	71076	88682
33066	92829	97349	50095	97538	55777	58994	22487	99974	03351
12453	15870	57421	79329	17487	03304	44107	25050	84297	75759
42973	12644	92911	56030	27487	73955	56921	33478	91622	26207
43397	07534	66071	65568	44354	19675	08492	81531	52961	95150
75824	23611	05961	23306	12030	99024	51409	09220	46436	25592
24270	90235	66887	91822	57484	20615	81048	06121	22790	30293
32155	55651	50165	40420	87805	94167	47014	03412	58232	01190
49128	68603	70371	73609	85851	27406	97846	10701	92339	75438
47881	21969	79326	14985	78919	70848	55693	15446	28657	08951
93028	12394	31791	66834	14037	59579	96851	03082	07339	70364
26557	21485	75834	65133	40810	62393	44524	88053	88774	22159
95069	64989	22449	98023	87914	58086	11783	21285	45201	65647
51473	69229	92738	22769	14238	94509	42403	66017	96637	19124
98957	80805	46441	33404	51741	28407	68943	08679	16198	15385
87592	01776	43166	78814	34823	73796	68837	75250	10046	10791
84100	16277	41706	72797	06683	16386	77954	53505	53812	78249
69462	24705	80114	02004	55594	08060	47933	63441	69719	09729
93608	75353	64552	45577	88199	58313	12097	63335	06115	16543
37241	51581	55513	98131	71564	23230	66610	06202	79080	36601
06855	76107	80948	61777	88260	44776	46862	15308	61721	61179
33864	01654	05023	13244	33936	93647	20956	30452	14649	74309
45167	23633	72357	55846	90477	73288	92447	79028	36627	20670
10376	76848	45729	44801	98129	64002	41643	11779	87581	70343
76057	26051	60638	94221	44527	99577	97790	44854	70527	18809
48157	59061	05958	68478	05364	39412	04714	51510	50286	81665
19216	87580	99632	26032	19660	21029	94943	14525	10052	81280
79799	94944	63412	38049	70926	37288	14214	59185	61717	97117
66450	25680	24907	14967	16427	12253	36493	25117	51827	24131
49643	44121	78339	23361	96049	32878	24339	84647	19902	21751
64688	52256	57958	17785	53250	65071	35052	10527	60814	66955
83199	86497	83727	31070	56191	20996	47452	64613	32828	50908
48957	96114	11718	10502	07780	43378	33125	27659	42041	17867
31520	90234	63855	30387	33228	22565	86551	90405	64928	35050
79921	55566	73325	84683	81471	77280	05845	56210	31429	92431
61925	86171	34231	66531	77294	69358	54647	96733	59454	22476
72272	85835	76108	32805	99814	70078	86787	06660	33271	06007
10468	53992	14394	29949	50095	84011	27467	89068	41882	89295

Table 1. *Continued.*

62553	31050	81817	50871	43793	11283	49570	55558	05747	46890
35447	95380	11951	12916	34031	76096	63437	93967	60777	42642
19103	27155	01880	79915	52008	63297	47031	29449	92174	31646
30337	62493	42159	33595	18578	17139	14653	05651	19936	74640
12368	37150	93911	77900	07710	44900	63036	09541	87084	70367
33570	88254	10300	95538	98565	40765	37423	39175	74793	16095
52449	30644	15956	64412	55242	70402	27976	08474	54064	17177
73618	42175	28226	68731	34494	83981	68435	13824	22642	85836
50313	70672	33804	69796	09370	97272	36512	37197	79443	68344
03759	11496	52998	41981	61478	56898	16524	99470	81524	78416
43816	09467	87381	69089	03858	55216	48997	44554	16474	62920
80437	07966	42383	76849	95261	64131	56586	93570	30307	65755
14066	48327	42374	32218	25867	36283	96019	91864	89841	50048
31114	11873	48718	83532	08240	96875	44598	90183	58624	51209
87913	07045	90226	53779	98807	84334	71407	75478	86113	25203
99132	12065	93418	08473	37561	73214	50471	47693	79254	09753
97629	31824	60163	55879	29385	79623	21296	64812	91567	82636
08978	72131	78267	03037	92041	89849	68399	91036	49158	98049
26668	13492	49538	04136	44762	29368	13949	77645	54521	62896
54846	41553	88909	84887	83911	17358	79433	98950	26714	33734
72087	63890	49539	35081	99740	97436	29133	23521	99344	35706
65497	79878	10520	11005	75748	94319	02478	72131	97000	56085
95803	73813	71494	87266	27061	52087	53429	04310	07814	50188
40865	42512	47260	19632	58445	89434	79864	73372	31074	65604
04433	29545	35821	07371	65764	91799	98243	52226	31903	79077
02057	08958	81921	22160	54557	91189	97243	52332	76120	41564
05041	35418	76418	43272	43076	34316	21812	98938	06714	09484
43891	14181	24985	29895	21869	01045	71527	20064	55337	97291
10374	41917	24948	25856	86498	46629	96251	41806	78913	18636
32879	79602	49067	60820	08714	92253	82848	73409	48597	34394
86673	82907	67765	73961	66363	61262	38162	31243	18387	11775
29063	82178	45025	54215	10231	28407	71873	58663	25027	97921
54587	40236	85404	66748	30574	95912	89247	74995	37696	73460
48607	87689	11871	98132	61211	32425	62083	99140	17050	57359
16094	97671	02064	35310	48094	23033	92444	40069	62889	29614
63697	17999	81953	97997	58143	55029	40358	52536	21476	68069
13117	55809	88704	52420	31357	59400	50199	33963	15282	12459
77399	52580	05822	09809	20640	47579	56527	83490	30383	51673
76618	27999	79590	15016	94053	10365	60327	50400	84668	75029
32875	25165	79676	05502	90404	32841	93419	72246	59709	65307
27458	54831	40982	14291	01684	19623	02560	37877	17419	23878
58754	72564	55632	06415	58533	69342	67019	16555	39796	07811
80475	59074	95621	92668	30545	48770	18343	64267	67114	85963
83370	87361	36193	46322	08986	50128	96736	52654	62464	84932
58667	32519	54144	10160	57730	78138	79983	91235	21796	61710
72694	94654	53848	76727	91635	81324	80402	89686	14023	66006
54230	03232	69368	30694	91077	07709	43411	54098	27967	06669
93447	01796	87049	02472	51265	20130	78615	59145	12773	61529
73087	46442	16168	64092	55380	39620	56090	28236	20743	46986
90745	38867	06363	80949	62878	76653	32971	27592	30049	12427
65291	78320	73014	50550	23378	95816	01401	81341	19325	16530
87733	37580	60372	32473	83102	66290	59967	32447	84792	54735
51120	84671	75765	89097	89408	43351	39652	19391	02850	72261
12974	05910	82732	72030	61112	91125	66991	20928	77852	05238
12779	27311	21722	01344	32040	15520	25040	86340	27990	33335

Table 2. Distribution of Student's t.

	Probability of a larger value of t, sign ignored				
df	0.5	0.1	0.05	0.01	0.001
1	1.00	6.31	12.71	63.66	636.62
2	0.82	2.92	4.30	9.92	31.60
3	0.76	2.35	3.18	5.84	12.92
4	0.74	2.13	2.78	4.60	8.61
5	0.73	2.02	2.57	4.03	6.87
6	0.72	1.94	2.45	3.71	5.96
7	0.71	1.89	2.36	3.50	5.41
8	0.71	1.86	2.31	3.36	5.04
9	0.70	1.83	2.26	3.25	4.78
10	0.70	1.81	2.23	3.17	4.59
11	0.70	1.80	2.20	3.11	4.44
12	0.70	1.78	2.18	3.05	4.32
13	0.69	1.77	2.16	3.01	4.22
14	0.69	1.76	2.14	2.98	4.14
15	0.69	1.75	2.13	2.95	4.07
16	0.69	1.75	2.12	2.92	4.01
17	0.69	1.74	2.11	2.90	3.97
18	0.69	1.73	2.10	2.88	3.92
19	0.69	1.73	2.09	2.86	3.88
20	0.69	1.72	2.09	2.85	3.85
21	0.69	1.72	2.08	2.83	3.82
22	0.69	1.72	2.07	2.82	3.79
23	0.69	1.71	2.07	2.81	3.77
24	0.68	1.71	2.06	2.80	3.75
25	0.68	1.71	2.06	2.79	3.73
26	0.68	1.71	2.06	2.78	3.71
27	0.68	1.70	2.05	2.77	3.69
28	0.68	1.70	2.05	2.76	3.67
29	0.68	1.70	2.05	2.76	3.66
30	0.68	1.70	2.04	2.75	3.65
40	0.68	1.68	2.02	2.70	3.55
60	0.68	1.67	2.00	2.66	3.46
120	0.68	1.66	1.98	2.62	3.37
∞	0.67	1.64	1.96	2.58	3.29

Table 3. Confidence intervals (95 percent) for the binomial distribution.

Observed	n = 10		n = 15		n = 20		n = 30		n = 50		n = 100	
0	0	31	0	22	0	17	0	12	0	7	0	4
1	0	45	0	32	0	25	0	17	0	11	0	5
2	3	56	2	40	1	32	1	22	0	14	0	7
3	7	65	4	48	3	38	2	27	1	17	1	9
4	12	74	8	55	6	44	4	31	2	19	1	10
5	19	81	12	62	9	49	6	35	3	22	2	11
6	26	88	16	68	12	54	8	39	5	24	2	13
7	35	93	21	73	15	59	10	42	6	27	3	14
8	44	97	27	79	19	64	12	46	7	29	4	15
9	55	100	32	84	23	68	15	49	9	31	4	16
10	69	100	38	88	27	73	17	53	10	34	5	18
11			45	92	32	77	20	56	12	36	6	19
12			52	96	36	81	23	59	13	38	6	20
13			60	98	41	85	25	63	15	40	7	21
14			68	100	46	88	28	66	16	42	8	22
15			78	100	51	91	31	69	18	45	9	24
16					56	94	34	72	20	47	9	25
17					62	97	37	75	21	49	10	26
18					68	99	41	77	23	51	11	27
19					75	100	44	80	25	53	12	28
20					83	100	47	83	26	55	13	29
21							51	85	28	57	13	30
22							54	88	30	59	14	31
23							58	90	32	61	15	32
24							61	92	34	63	16	34
25							65	94	36	64	17	35
26							69	96	37	66	18	36
27							73	98	39	68	19	37
28							78	99	41	70	19	38
29							83	100	43	72	20	39
30							88	100	45	74	21	40
31									47	75	22	41
32									49	77	23	42
33									51	79	24	43
34									53	80	25	44
35									55	82	26	45
36									58	84	27	46
37									60	85	28	47
38									62	87	28	48
39									64	88	29	49
40									66	90	30	50
41									69	91	31	51
42									71	93	32	52
43									73	94	33	53
44									76	95	34	54
45									78	97	35	55
46									81	98	36	56
47									83	99	37	57
48									86	100	38	58
49									89	100	39	59
50									93	100	40	60

Table 4. The ArcSine $\sqrt{\text{Percentage}}$ transformation. Transformation of binomial percentages in the margins to angles equals information in degrees.

%	0	1	2	3	4	5	6	7	8	9
0	0.00	0.57	0.81	0.99	1.15	1.28	1.40	1.52	1.62	1.72
0.1	1.81	1.90	1.99	2.07	2.14	2.22	2.29	2.36	2.43	2.50
0.2	2.56	2.63	2.69	2.75	2.81	2.87	2.92	2.98	3.03	3.09
0.3	3.14	3.19	3.24	3.29	3.34	3.39	3.44	3.49	3.53	3.58
0.4	3.63	3.67	3.72	3.76	3.80	3.85	3.89	3.93	3.97	4.01
0.5	4.05	4.10	4.14	4.17	4.21	4.25	4.29	4.33	4.37	4.41
0.6	4.44	4.48	4.52	4.55	4.59	4.62	4.66	4.70	4.73	4.76
0.7	4.80	4.83	4.87	4.90	4.93	4.97	5.00	5.03	5.07	5.10
0.8	5.13	5.16	5.20	5.23	5.26	5.29	5.32	5.35	5.38	5.41
0.9	5.44	5.47	5.50	5.53	5.56	5.59	5.62	5.65	5.68	5.71
1	5.74	6.02	6.29	6.55	6.80	7.03	7.27	7.49	7.71	7.92
2	8.13	8.33	8.53	8.72	8.91	9.10	9.28	9.46	9.63	9.80
3	9.97	10.14	10.30	10.47	10.63	10.78	10.94	11.09	11.24	11.39
4	11.54	11.68	11.83	11.97	12.11	12.25	12.38	12.52	12.66	12.79
5	12.92	13.05	13.18	13.31	13.44	13.56	13.69	13.81	13.94	14.06
6	14.18	14.30	14.42	14.54	14.65	14.77	14.89	15.00	15.12	15.23
7	15.34	15.45	15.56	15.68	15.79	15.89	16.00	16.11	16.22	16.32
8	16.43	16.54	16.64	16.74	16.85	16.95	17.05	17.16	17.26	17.36
9	17.46	17.56	17.66	17.76	17.85	17.95	18.05	18.15	18.24	18.34
10	18.43	18.53	18.63	18.72	18.81	18.91	19.00	19.09	19.19	19.28
11	19.37	19.46	19.55	19.64	19.73	19.82	19.91	20.00	20.09	20.18
12	20.27	20.36	20.44	20.53	20.62	20.70	20.79	20.88	20.96	21.05
13	21.13	21.22	21.30	21.39	21.47	21.56	21.64	21.72	21.81	21.89
14	21.97	22.06	22.14	22.22	22.30	22.38	22.46	22.54	22.63	22.71
15	22.79	22.87	22.95	23.03	23.11	23.18	23.26	23.34	23.42	23.50
16	23.58	23.66	23.73	23.81	23.89	23.97	24.04	24.12	24.20	24.27
17	24.35	24.43	24.50	24.58	24.65	24.73	24.80	24.88	24.95	25.03
18	25.10	25.18	25.25	25.33	25.40	25.47	25.55	25.62	25.70	25.77
19	25.84	25.91	25.99	26.06	26.13	26.21	26.28	26.35	26.42	26.49
20	26.57	26.64	26.71	26.78	26.85	26.92	26.99	27.06	27.13	27.20
21	27.27	27.35	27.42	27.49	27.56	27.62	27.69	27.76	27.83	27.90
22	27.97	28.04	28.11	28.18	28.25	28.32	28.39	28.45	28.52	28.59
23	28.66	28.73	28.79	28.86	28.93	29.00	29.06	29.13	29.20	29.27
24	29.33	29.40	29.47	29.53	29.60	29.67	29.73	29.80	29.87	29.93
25	30.00	30.07	30.13	30.20	30.26	30.33	30.40	30.46	30.53	30.59
26	30.66	30.72	30.79	30.85	30.92	30.98	31.05	31.11	31.18	31.24
27	31.31	31.37	31.44	31.50	31.56	31.63	31.69	31.76	31.82	31.88
28	31.95	32.01	32.08	32.14	32.20	32.27	32.33	32.39	32.46	32.52
29	32.58	32.65	32.71	32.77	32.83	32.90	32.96	33.02	33.09	33.15
30	33.21	33.27	33.34	33.40	33.46	33.52	33.58	33.65	33.71	33.77
31	33.83	33.90	33.96	34.02	34.08	34.14	34.20	34.27	34.33	34.39
32	34.45	34.51	34.57	34.63	34.70	34.76	34.82	34.88	34.94	35.00
33	35.06	35.12	35.18	35.24	35.30	35.37	35.43	35.49	35.55	35.61
34	35.67	35.73	35.79	35.85	35.91	35.97	36.03	36.09	36.15	36.21
35	36.27	36.33	36.39	36.45	36.51	36.57	36.63	36.69	36.75	36.81
36	36.87	36.93	36.99	37.05	37.11	37.17	37.23	37.29	37.35	37.41
37	37.46	37.52	37.58	37.64	37.70	37.76	37.82	37.88	37.94	38.00
38	38.06	38.12	38.17	38.23	38.29	38.35	38.41	38.47	38.53	38.59
39	38.65	38.70	38.76	38.82	38.88	38.94	39.00	39.06	39.11	39.17
40	39.23	39.29	39.35	39.41	39.47	39.52	39.58	39.64	39.70	39.76
41	39.82	39.87	39.93	39.99	40.05	40.11	40.16	40.22	40.28	40.34
42	40.40	40.45	40.51	40.57	40.63	40.69	40.74	40.80	40.86	40.92
43	40.98	41.03	41.09	41.15	41.21	41.27	41.32	41.38	41.44	41.50
44	41.55	41.61	41.67	41.73	41.78	41.84	41.90	41.96	42.02	42.07
45	42.13	42.19	42.25	42.30	42.36	42.42	42.48	42.53	42.59	42.65
46	42.71	42.76	42.82	42.88	42.94	42.99	43.05	43.11	43.17	43.22
47	43.28	43.34	43.39	43.45	43.51	43.57	43.62	43.68	43.74	43.80
48	43.85	43.91	43.97	44.03	44.08	44.14	44.20	44.26	44.31	44.37
49	44.43	44.48	44.54	44.60	44.66	44.71	44.77	44.83	44.89	44.94
50	45.00	45.06	45.11	45.17	45.23	45.29	45.34	45.40	45.46	45.52
51	45.57	45.63	45.69	45.74	45.80	45.86	45.92	45.97	46.03	46.09
52	46.15	46.20	46.26	46.32	46.38	46.43	46.49	46.55	46.61	46.66
53	46.72	46.78	46.83	46.89	46.95	47.01	47.06	47.12	47.18	47.24

Table 4. *Continued.*

%	0	1	2	3	4	5	6	7	8	9
54	47.29	47.35	47.41	47.47	47.52	47.58	47.64	47.70	47.75	47.81
55	47.87	47.93	47.98	48.04	48.10	48.16	48.22	48.27	48.33	48.39
56	48.45	48.50	48.56	48.62	48.68	48.73	48.79	48.85	48.91	48.97
57	49.02	49.08	49.14	49.20	49.26	49.31	49.37	49.43	49.49	49.55
58	49.60	49.66	49.72	49.78	49.84	49.89	49.95	50.01	50.07	50.13
59	50.18	50.24	50.30	50.36	50.42	50.48	50.53	50.59	50.65	50.71
60	50.77	50.83	50.89	50.94	51.00	51.06	51.12	51.18	51.24	51.30
61	51.35	51.41	51.47	51.53	51.59	51.65	51.71	51.77	51.83	51.88
62	51.94	52.00	52.06	52.12	52.18	52.24	52.30	52.36	52.42	52.48
63	52.54	52.59	52.65	52.71	52.77	52.83	52.89	52.95	53.01	53.07
64	53.13	53.19	53.25	53.31	53.37	53.43	53.49	53.55	53.61	53.67
65	53.73	53.79	53.85	53.91	53.97	54.03	54.09	54.15	54.21	54.27
66	54.33	54.39	54.45	54.51	54.57	54.63	54.70	54.76	54.82	54.88
67	54.94	55.00	55.06	55.12	55.18	55.24	55.30	55.37	55.43	55.49
68	55.55	55.61	55.67	55.73	55.80	55.86	55.92	55.98	56.04	56.10
69	56.17	56.23	56.29	56.35	56.42	56.48	56.54	56.60	56.66	56.73
70	56.79	56.85	56.91	56.98	57.04	57.10	57.17	57.23	57.29	57.35
71	57.42	57.48	57.54	57.61	57.67	57.73	57.80	57.86	57.92	57.99
72	58.05	58.12	58.18	58.24	58.31	58.37	58.44	58.50	58.56	58.63
73	58.69	58.76	58.82	58.89	58.95	59.02	59.08	59.15	59.21	59.28
74	59.34	59.41	59.47	59.54	59.60	59.67	59.74	59.80	59.87	59.93
75	60.00	60.07	60.13	60.20	60.27	60.33	60.40	60.47	60.53	60.60
76	60.67	60.73	60.80	60.87	60.94	61.00	61.07	61.14	61.21	61.27
77	61.34	61.41	61.48	61.55	61.61	61.68	61.75	61.82	61.89	61.96
78	62.03	62.10	62.17	62.24	62.31	62.38	62.44	62.51	62.58	62.66
79	62.73	62.80	62.87	62.94	63.01	63.08	63.15	63.22	63.29	63.36
80	63.43	63.51	63.58	63.65	63.72	63.79	63.87	63.94	64.01	64.09
81	64.16	64.23	64.30	64.38	64.45	64.53	64.60	64.67	64.75	64.82
82	64.90	64.97	65.05	65.12	65.20	65.27	65.35	65.42	65.50	65.57
83	65.65	65.73	65.80	65.88	65.96	66.03	66.11	66.19	66.27	66.34
84	66.42	66.50	66.58	66.66	66.74	66.82	66.89	66.97	67.05	67.13
85	67.21	67.29	67.37	67.46	67.54	67.62	67.70	67.78	67.86	67.94
86	68.03	68.11	68.19	68.28	68.36	68.44	68.53	68.61	68.70	68.78
87	68.87	68.95	69.04	69.12	69.21	69.30	69.38	69.47	69.56	69.64
88	69.73	69.82	69.91	70.00	70.09	70.18	70.27	70.36	70.45	70.54
89	70.63	70.72	70.81	70.91	71.00	71.09	71.19	71.28	71.37	71.47
90	71.57	71.66	71.76	71.85	71.95	72.05	72.15	72.24	72.34	72.44
91	72.54	72.64	72.74	72.85	72.95	73.05	73.15	73.26	73.36	73.46
92	73.57	73.68	73.78	73.89	74.00	74.11	74.21	74.32	74.44	74.55
93	74.66	74.77	74.88	75.00	75.11	75.23	75.35	75.46	75.58	75.70
94	75.82	75.94	76.06	76.19	76.31	76.44	76.56	76.69	76.82	76.95
95	77.08	77.21	77.34	77.48	77.62	77.75	77.89	78.03	78.17	78.32
96	78.46	78.61	78.76	78.91	79.06	79.22	79.37	79.53	79.70	79.86
97	80.03	80.20	80.37	80.54	80.72	80.90	81.09	81.28	81.47	81.67
98	81.87	82.08	82.29	82.51	82.73	82.97	83.20	83.45	83.71	83.98
99	84.26	84.29	84.32	84.35	84.38	84.41	84.44	84.47	84.50	84.53
99.1	84.56	84.59	84.62	84.65	84.68	84.71	84.74	84.77	84.80	84.84
99.2	84.87	84.90	84.93	84.97	85.00	85.03	85.07	85.10	85.13	85.17
99.3	85.20	85.24	85.27	85.30	85.34	85.38	85.41	85.45	85.48	85.52
99.4	85.56	85.59	85.63	85.67	85.71	85.75	85.79	85.83	85.86	85.90
99.5	85.95	85.99	86.03	86.07	86.11	86.15	86.20	86.24	86.28	86.33
99.6	86.37	86.42	86.47	86.51	86.56	86.61	86.66	86.71	86.76	86.81
99.7	86.86	86.91	86.97	87.02	87.08	87.13	87.19	87.25	87.31	87.37
99.8	87.44	87.50	87.57	87.64	87.71	87.78	87.86	87.93	88.01	88.10
99.9	88.19	88.28	88.38	88.48	88.60	88.72	88.85	89.01	89.19	89.43

Table 5. Two-tailed significance levels of the correlation coefficient *r*.

	Significance level			
df	**0.1**	**0.05**	**0.01**	**0.001**
1	0.988	0.997	1.000	1.000
2	0.900	0.950	0.990	0.999
3	0.805	0.878	0.959	0.991
4	0.729	0.811	0.917	0.974
5	0.669	0.754	0.875	0.951
6	0.621	0.707	0.834	0.925
7	0.582	0.666	0.798	0.898
8	0.549	0.632	0.765	0.872
9	0.521	0.602	0.735	0.847
10	0.497	0.576	0.708	0.823
11	0.476	0.553	0.684	0.801
12	0.458	0.532	0.661	0.780
13	0.441	0.514	0.641	0.760
14	0.426	0.497	0.623	0.742
15	0.412	0.482	0.606	0.725
16	0.400	0.468	0.590	0.708
17	0.389	0.456	0.575	0.693
18	0.378	0.444	0.561	0.679
19	0.369	0.433	0.549	0.665
20	0.360	0.423	0.537	0.652
21	0.352	0.413	0.526	0.640
22	0.344	0.404	0.515	0.629
23	0.337	0.396	0.505	0.618
24	0.330	0.388	0.496	0.607
25	0.323	0.381	0.487	0.597
26	0.317	0.374	0.479	0.588
27	0.311	0.367	0.471	0.579
28	0.306	0.361	0.463	0.570
29	0.301	0.355	0.456	0.562
30	0.296	0.349	0.449	0.554
40	0.257	0.304	0.393	0.490
50	0.231	0.273	0.354	0.443
60	0.211	0.250	0.325	0.408
70	0.195	0.232	0.302	0.380
80	0.183	0.217	0.283	0.357
90	0.173	0.205	0.267	0.338
100	0.164	0.195	0.254	0.321
150	0.134	0.159	0.208	0.264
200	0.116	0.138	0.181	0.230
300	0.095	0.113	0.148	0.188
400	0.082	0.098	0.128	0.164
500	0.073	0.088	0.115	0.146

Appendix 4. Statistical Analysis Worked Examples

The best way to understand statistics is to work through many examples. Unfortunately, most examples are computationally intensive. But with the computing power available on the desktop, there are many good statistical computing packages available. In this Appendix, we show the commands as well as output for many of the examples presented throughout. We choose the R language because it is freely available and because it provides an extensive array of statistical analysis procedures. The R package can be downloaded from the R project web site at http://www.r-project.org/.

The application of the methods discussed throughout this book is computationally intensive. We present examples of these analyses in this section; these worked examples may be used as starting points or templates for other analyses. Included in this Appendix are descriptions of the data sets as well as a variety of sampling methods with results. The data sets as well as the programs can be downloaded from RMRS http://www.fs.fed.us/rm/ftcol/index.shtm.

File Name	Description
schreuderworkedexamples.xls	Original large data set with description and summary of results
surinam.csv	Text export of large data set that is read by R
macros.r	Miscellaneous function definitions used by the R programs
schreuder.r	Calculations used in the text body
schreudertables.r	Development of tables used in appendices
workedexamples.r	Worked examples using the large data set

Analysis Software

The choice of software running on Windows, Linux, or other platforms is very broad. The commercially available packages such as SAS, SPSS, or S/S-Plus run on the full range of platforms, from PC to mainframe. Because of the uniqueness and selective availability of each of these packages, we do not attempt to work these examples in terms of these systems. Instead, we illustrate the analyses with the readily available open source package, R. The R data handling package is extremely robust and powerful, and it offers a wide array of statistical analysis procedures; most programs written for the widely available commercial packages S and S-Plus will run under the R system. Links from the R home page will take you to the downloads for the package itself (the complete installation contains the executables along with complete documentation), as well as contributed packages and various electronic publications in English, Spanish, French, and German.

Data Sets

The first data set is a small data set of 10 trees that is presented in Table 1 in the text body. Although contrived, it is an easy data set to analyze by hand.

The second data set consists of a 60 ha stem-mapped population of trees from a tropical forest in Surinam. These data were used and described by Schreuder and others (1997). The tree heights and volumes were added by using trees of the same size from FIA data for very different species by necessity. This population of 6,806 trees has the relative spatial location of the trees and is used to illustrate the efficiency of several sampling strategies. The population stem map is displayed in Figure A-2 below.

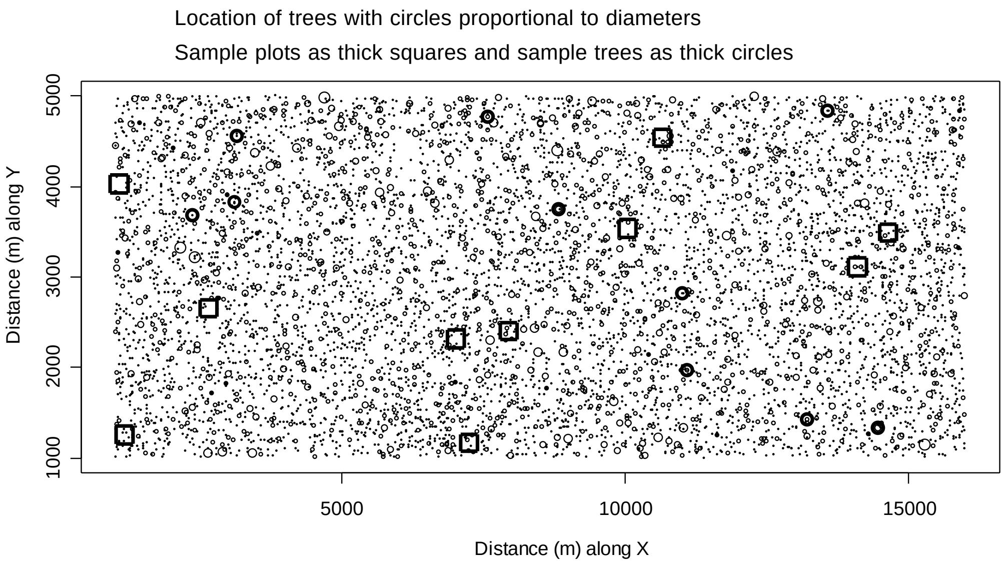


Figure A-2. Stem map for the Surinam population with sample locations for a SRS and a cluster sample.

The attributes recorded for each tree were:

Column names	**Description**
Diameter_cm	Diameter of tree measured in centimeters
Longitude	X-offset of tree measured in 0.1m
Lattitude	Y-offset of tree measured in 0.1m
Height_m	Height of the tree in m
Volume_cum	Volume of the tree in m^3
Subplot	Subplot identification based on grid labeled with letters for one dimension and numbers for the other
DBHClass	Diameter class
Diameter_in	Diameter of tree measured in inches - hard conversion
Height_ft	Height of the tree in feet - hard conversion
Volume_cuft	Volume of the tree in cubic ft - hard conversion
CC	Crown class of (D)ominant or (S)ubdominant derived from height

The tree locations are indicated with circles that are proportional to the diameter of the tree. Ten trees were selected at random from the population to illustrate simple random sampling; these trees are indicated with thick circles. Ten 30-m by 30-m plots were also randomly selected to illustrate cluster sampling; these plots are indicated with thick squares. Stratifed sampling is illustrated by categorizing the trees as either dominants or subdominants on the basis of height.

Results

The results from analyzing the small data set are tabulated in the main body text. The worked example calculations can be created by running the R program in the file *schreuder.r*.

The Surinam data set can be used for realistic exercises in applying the methods discussed in this book. One of the most useful steps in any analysis is to produce some descriptive statistics, either tabular or graphical. Some useful graphics include the boxplot. Examples of the boxplot for the volume (in m^3) for the entire population and the boxplot for the stratified population are shown in Figure A-3.

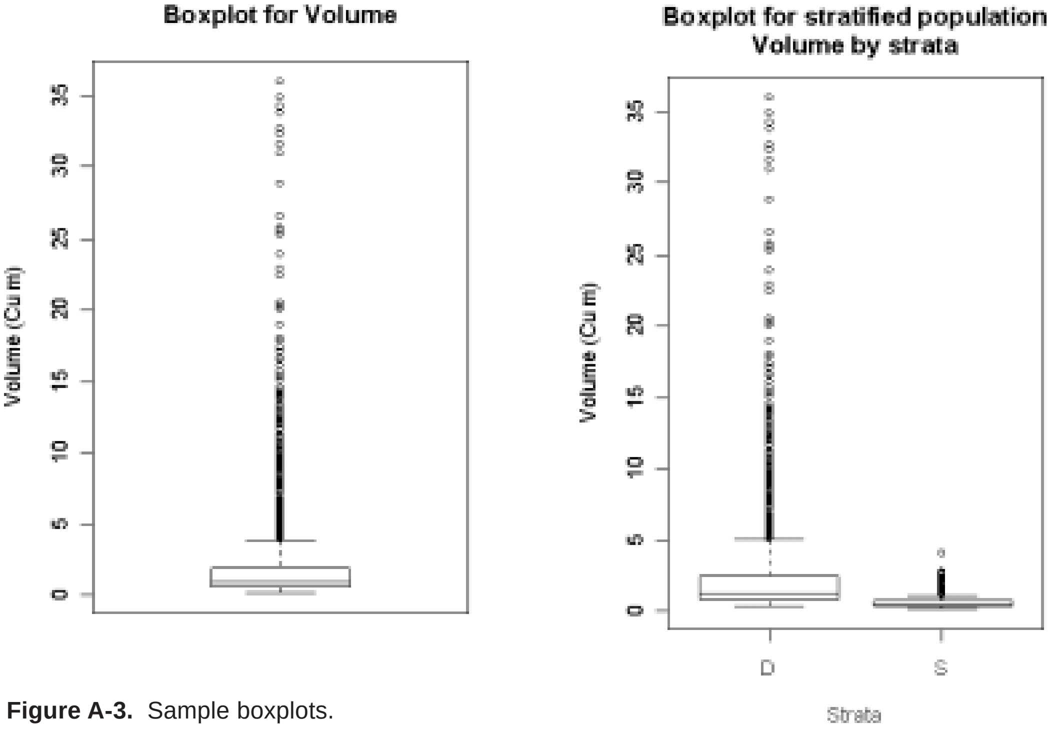


Figure A-3. Sample boxplots.

Clearly, the mean volume per tree is related to the crown position of the tree, and thus would be useful in stratifying the population.

The sampling methods discussed in this paper all have strengths and weaknesses. As an exercise, the original population and five samples were used to demonstrate actual calculations performed in the R analysis program. The commands to duplicate these results are in the R command file named *workedexamples.r*

Appendix 3, Table 6 summarizes some of the results.

The first row of the table contains the population parameters calculated from all 6,806 trees; this is the benchmark against which all estimates are judged. The next two lines of the table are for two contrived samples that picked three small trees and then three large trees. Even though this is a woefully inadequate sample size, either could actually result from a random trial. Both result in poor estimates of the population parameter. The nature of estimates resulting from a random draw do not guarantee reliable estimates. The next trial was a SRS of size 10 selected from this population. The estimate, again sample based, is better, but still not reliable. A random sample of 10 clusters resulted in the measurement of 73 trees and yielded a reliable result. A stratified sample measured 10 trees, but this time, five from the dominant and five from the subdominant classes. This particular trial resulted in another reliable estimate of the population parameter.

Table 6. Summary results for population parameters of Surinam population and results for some samples.

Method	Size	Mean	Variance
Original population	6806	1.89	8.30
Sample of small trees	3	0.40	0.004
Sample of large trees	3	8.44	10.54
Simple random sample	10	4.15	14.94
Cluster sample	10 plots 73 trees	2.25	3.17
Stratified sample	10	2.35	7.29

The output resulting from running the large data set analysis file follow:

```
************************************************************
Worked examples:
Surinam data set.
Summary information for the population.
The following variables are available:
 Diameter.cm Longitude Lattitude Height.m Volume.cum SubPolt x2 x3 DBHClass
Diameter.in Height.ft Volume.cuft CC
The basic pop statistics for these variables are:
Statistical summary:
         Diameter.cm   Height.m  Volume.cum
Mean        41.79327   24.28212    1.886360
Variance   326.25685   71.81564    8.295587
N         6806.00000 6806.00000 6806.000000
The statistics are saved to file:  ..//Data/AllSurinamResults.csv
With the distribution summary:
  Diameter.cm          Height.m         Volume.cum
 “Min.  : 25.00  ““Min.  : 9.20  ““Min.  : 0.176  “
 “1st Qu.: 29.00  ““1st Qu.:19.50  ““1st Qu.: 0.569  “
 “Median : 36.00  ““Median :22.60  ““Median : 0.961  “
 “Mean   : 41.79  ““Mean   :24.28  ““Mean   : 1.886  “
 “3rd Qu.: 48.00  ““3rd Qu.:25.90  ““3rd Qu.: 1.903  “
 “Max.  :165.00  ““Max.  :70.90  ““Max.  :35.978  “
 NA                 NA                 NA
  ************************************************************
See plot (in another window) for spatial arrangement of trees.
In addition, type:
  identify(surinam$Longitude, surinam$Lattitude,surinam$Diameter.cm)
to interactively click on points to identify diameter.
NOTE: be sure to rt-click-Stop if you do this.
See plot (in another window) example of boxplot for diameter.
  ************************************************************
Suppose we select a SRS of size three from the population, say observations:
 4 34 216
The basic sample statistics for this sample are:
Statistical summary:
         Diameter.cm Height.m Volume.cum
Mean              88 42.66667   8.444333
Variance          39 90.20333  10.535722
n                  3  3.00000   3.000000
The statistics will be saved to file:  ..//Data/SurinamSample1Results.csv
  ************************************************************
Suppose we select another SRS of size three from the population, say observations:
 3 814 1278
The basic sample statistics for this sample are:
Statistical summary:
         Diameter.cm Height.m  Volume.cum
Mean              28    18.10 0.402333333
Variance          12    50.89 0.004058333
n                  3     3.00 3.000000000
The statistics will be saved to file://Data/SurinamSample2Results.csv
  ************************************************************
Compare the estimates from these two samples with the actual population parameters.
  ************************************************************
Now let us select a true random sample of size 10, say observations:
 654 1008 1040 3038 3587 4529 4564 5470 5628 6030
The basic sample statistics for this sample are:
Statistical summary:
         Diameter.cm Height.m Volume.cum
```

```
Mean         57.3000  32.9100    4.15060
Variance    683.1222 238.1877   14.93925
n            10.0000  10.0000   10.00000

  ************************************************************
How did we do in estimating the population parameters?
  ************************************************************

  ************************************************************
We could also use plots or clusters to sample.
Suppose we select a true random sample of 10 plots

The plot samples selected are:
     Diameter.cm Height.m Volume.cum IsOnPlot
37            35     25.0      1.196        2
46            85     32.3      3.761        2
75            32     19.2      0.710        8
77            28     20.1      0.544        8
78            64     46.9      5.116        8
79            28     20.1      0.545        8
80            30     28.7      0.734        8
81            28     22.3      0.677        8
82            46     20.1      1.403        8
84            46     14.9      0.927        8
85            26     22.3      0.585        8
86            40     23.5      1.220        8
87            27     21.3      0.648        8
88            25     17.5      0.375        8
89            32     20.1      0.743        8
90            32     26.8      0.936        8
763           57     21.9      2.520        7
767           31     23.8      0.737        7
770           41     29.3      1.547        7
776           47     23.5      1.834        7
2765          42     25.0      1.358        9
2766          25     21.0      0.397        9
2768          57     18.6      1.731        9
2771          32     25.0      0.874        9
2775          35     19.8      0.942        9
2776          27     12.2      0.286        9
2778          39     26.5      1.311        9
2779          42     24.1      1.540        9
2780          42     22.6      1.281        9
2878          44     20.7      1.205        6
2879          29     19.5      0.515        6
2880          33     22.3      0.816        6
2881          36     23.2      1.033        6
2882          71     32.3      2.962        6
2883          82     47.2      7.719        6
2885          25     20.4      0.373        6
2886          29     22.3      0.584        6
2887          60     28.4      3.769        6
2888          26     22.6      0.555        6
2889          25     17.4      0.390        6
2890          33     21.0      0.780        6
2891          40     17.7      1.049        6
2892          72     42.5      4.457        6
3191          26     14.6      0.295        3
3192          65     44.2      5.306        3
3193          29     14.6      0.414        3
3196          32     19.1      0.658        3
4116          28     18.3      0.480        1
4120          89     40.3      7.985        1
4121          29     17.7      0.505        1
4123          52     19.9      1.676        1
4124          25     23.5      0.487        1
4127          26     15.8      0.407        1
4129          72     27.4      5.467        1
4386          43     25.0      1.359        4
4393          27     22.6      0.697        4
```

```
5863          30     18.9      0.541        5
5864          36     21.3      0.920        5
5866          33     24.7      0.721        5
5867         117     63.6     22.413        5
5868          65     48.9      6.584        5
5872          31     21.0      0.732        5
5874         120     67.6     25.699        5
5875          36     27.7      1.302        5
5876          31     21.4      0.717        5
5877          46     23.2      1.701        5
5878          29     19.5      0.571        5
5879          38     19.8      0.973        5
6118          28     22.3      0.617       10
6119          38     12.8      0.450       10
6120          26     21.6      0.485       10
6123          43     25.0      1.359       10
6125          96     46.3      9.713       10
Cluster Statistics for the volume (CuM):
          [1]      [2]      [3]      [4]      [5]       [6]       [7]
mi    7.000000 2.000000 4.000000 2.000000 12.00000 14.000000 4.0000000
ybari 2.429571 2.478500 1.668250 1.028000  5.23950  1.871929 1.6595000
vari  9.337635 3.289613 5.904263 0.219122 80.47928  4.574484 0.5448577
           [8]       [9]     [10]
mi    14.000000 9.0000000  5.00000
ybari  1.083071 1.0800000  2.52480
vari   1.422464 0.2456065 16.28365
Thus the estimates of the volume for the total population are:
  Mean:  2.245466  and variance:  3.167461
  ***********************************************************
  ***********************************************************
Assume we stratify the population by crown class.
Stratified Surinam data set.
Summary info for the stratified population.
The following variables are available:
 Diameter.cm Longitude Lattitude Height.m Volume.cum SubPolt x2 x3 DBHClass
Diameter.in Height.ft Volume.cuft CC IsSRS IsOnPlot IsStratifiedSample
The basic population statistics for the dominants are:
Statistical summary:
         Diameter.cm   Height.m  Volume.cum
Mean        45.86105   27.10766    2.400007
Variance   377.56259   71.87200   10.712216
N         4829.00000 4829.00000 4829.000000
The statistics will be saved to file://Data/DomSurinamResults.csv
While the statistics for the suppressed trees are:
Statistical summary:
         Diameter.cm    Height.m   Volume.cum
Mean        31.85736   17.380475    0.6317299
Variance    61.79857    4.544404    0.1742217
N         1977.00000 1977.000000 1977.0000000
The statistics will be saved to file://Data/SupSurinamResults.csv
  ***********************************************************
Compare the parameters from these two strata with the single population parameters.
  ***********************************************************
We can sample from these two strata with the results:
Stratified Stats for the volume (CuM):
      [1]         [2]
IDi   “D”         “S”
nh    “5”         “5”
ybarh “3.053”     “0.6374”
varh  “10.1608805” “0.3070853”

Resulting in population estimates of:
   Mean:  2.351319  and variance:  7.29087
See plot (in another window) example of boxplot for diameter, by dominance.
```

Index

The use of trade and company names is for the benefit of the reader; such use does not constitute an official endorsement or approval of any service or product by the U.S. Department of Agriculture to the exclusion of others that may be suitable.

RMRS

ROCKY MOUNTAIN RESEARCH STATION

The Rocky Mountain Research Station develops scientific information and technology to improve management, protection, and use of the forests and rangelands. Research is designed to meet the needs of the National Forest managers, Federal and State agencies, public and private organizations, academic institutions, industry, and individuals.

Studies accelerate solutions to problems involving ecosystems, range, forests, water, recreation, fire, resource inventory, land reclamation, community sustainability, forest engineering technology, multiple use economics, wildlife and fish habitat, and forest insects and diseases. Studies are conducted cooperatively, and applications may be found worldwide.

Research Locations

Flagstaff, Arizona
Fort Collins, Colorado*
Boise, Idaho
Moscow, Idaho
Bozeman, Montana
Missoula, Montana
Lincoln, Nebraska
Reno, Nevada
Albuquerque, New Mexico
Rapid City, South Dakota
Logan, Utah
Ogden, Utah
Provo, Utah
Laramie, Wyoming

*Station Headquarters, Natural Resources Research Center, 2150 Centre Avenue, Building A, Fort Collins, CO 80526.

The U.S. Department of Agriculture (USDA) prohibits discrimination in all its programs and activities on the basis of race, color, national origin, sex, religion, age, disability, political beliefs, sexual orientation, or marital or family status. (Not all prohibited bases apply to all programs.) Persons with disabilities who require alternative means for communication of program information (Braille, large print, audio-tape, etc.) should contact USDA's TARGET Center at (202) 720-2600 (voice and TDD).

To file a complaint of discrimination, write USDA, Director, Office of Civil Rights, Room 326 W, Whitten Building, 1400 Independence Avenue, SW, Washington, D.C. 20250-9410 or call (202) 720-5964 (voice and TDD). USDA is an equal opportunity provider and employer.

Federal Recycling Program Printed on Recycled Paper