ELEMENTARY FOREST SAMPLING

Agriculture Handbook No. 232

U.S. Department of Agriculture • Forest Service

ELEMENTARY FOREST SAMPLING

FRANK FREESE
Southern Forest Experiment Station, Forest Service

Agriculture Handbook No. 232 December 1962

U.S. Department of Agriculture • Forest Service

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington 25, D.C. - Price 50 cents

ACKNOWLEDGMENTS

I should like to express my appreciation to Professor George W. Snedecor of the Iowa State University Statistical Laboratory and to the Iowa State University Press for their generous permission to reprint tables 1, 3, and 4 from their book *Statistical Methods*, 5th edition. Thanks are also due to Dr. C. I. Bliss of the Connecticut Agricultural Experiment Station, who originally prepared the material in table 4. I am indebted to Professor Sir Ronald A. Fisher, F.R.S., Cambridge, and to Dr. Frank Yates, F.R.S., Rothamsted, and to Messrs. Oliver and Boyd Ltd., Edinburgh, for permission to reprint table 2 from their book *Statistical Tables for Biological, Agricultural, and Medical Research.*

FRANK FREESE
Southern Forest Experiment Station

CONTENTS

ELEMENTARY FOREST SAMPLING

This is a statistical cookbook for foresters. It presents some sampling methods that have been found useful in forestry. No attempt is made to go into the theory behind these methods. This has some dangers, but experience has shown that few foresters will venture into the intricacies of statistical theory until they are familiar with some of the common sampling designs and computations.

The aim here is to provide that familiarity. Readers who attain such familiarity will be able to handle many of the routine sampling problems. They will also find that many problems have been left unanswered and many ramifications of sampling ignored. It is hoped that when they reach this stage they will delve into more comprehensive works on sampling. Several very good ones are listed on page 78.

BASIC CONCEPTS

Why Sample?

Most human decisions are made with incomplete knowledge. In daily life, a physician may diagnose disease from a single drop of blood or a microscopic section of tissue; a housewife judges a watermelon by its "plug" or by the sound it emits when thumped; and amid a bewildering array of choices and claims we select toothpaste, insurance, vacation spots, mates, and careers with but a fragment of the total information necessary or desirable for complete understanding. All of these we do with the ardent hope that the drop of blood, the melon plug, and the advertising claim give a reliable picture of the population they represent.

In manufacturing and business, in science, and no less in forestry, partial knowledge is a normal state. The complete census is rare—the sample is commonplace. A ranger must advertise timber sales with estimated volume, estimated grade yield and value, estimated cost, and estimated risk. The nurseryman sows seed whose germination is estimated from a tiny fraction of the seedlot, and at harvest he estimates the seedling crop with sample counts in the nursery beds. Enterprising pulp companies, seeking a source of raw material in sawmill residue, may estimate the potential tonnage of chippable material by multiplying reported production by a set of conversion factors obtained at a few representative sawmills.

However desirable a complete measurement may seem, there are several good reasons why sampling is often preferred. In the first place, complete measurement or enumeration may be impossible. The nurseryman might be somewhat better informed if he knew the germinative capacity of all the seed to be sown, but the destructive nature of the germination test precludes testing every seed. For identical reasons, it is impossible to measure the bending strength of all the timbers to be used in a bridge, the tearing strength of all the paper to be put into a book, or the grade of all the boards to be produced in a timber sale. If the tests were permitted, no seedlings would be produced, no bridges would be built, no books printed, and no stumpage sold. Clearly where testing is destructive, some sort of sampling is inescapable.

In other instances total measurement or count is not feasible. Consider the staggering task of testing the quality of all the water in a reservoir, weighing all the fish in a stream, counting all the seedlings in a 500-bed nursery, enumerating all the egg masses in a turpentine beetle infestation, measuring diameter and height of all the merchantable trees in a 10,000-acre forest. Obviously, the enormity of the task would demand some sort of sampling procedure.

It is well known that sampling will frequently provide the essential information at a far lower cost than a complete enumeration. Less well known is the fact that this information may at times be more reliable than that obtained by a 100-percent inventory. There are several reasons why this might be true. With fewer observations to be made and more time available, measurement of the units in the sample can be and is more likely to be made with greater care. In addition, a portion of the saving resulting from sampling could be used to buy better instruments and to employ or train higher caliber personnel. It is not hard to see that good measurements on 5 percent of the units in a population could provide more reliable information than sloppy measurements on 100 percent of the units.

Finally, since sample data can be collected and processed in a fraction of the time required for a complete inventory, the information obtained may be more timely. Surveying 100 percent of the lumber market is not going to provide information that is very useful to a seller if it takes 10 months to complete the job.

Populations, Parameters, and Estimates

The central notion in any sampling problem is the existence of a population. It is helpful to think of a population as an aggregate of unit values, where the "unit" is the thing upon which the observation is made, and the "value" is the property observed on that thing. For example, we may imagine a square 40-acre 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 aggregate of all heights of trees on the specified forty. The diameters of these same trees would be another population. The cubic volumes in some particular portion of the stems constitute still another population.

Alternatively, the units might be defined as the 400 1-chain-square plots into which the tract could be divided. The cubic volumes of trees on these plots might form one population. The board-foot volumes of the same trees would be another population. The number of earthworms in the top 6 inches of soil on these plots could be still a third population.

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 6 inches 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 both be very difficult.

To characterize the population as a whole, we often use certain constants that are called parameters. The mean value per plot in a population of quarter-acre plots is a parameter. The proportion of living seedlings in a pine plantation is a parameter. The total number of units in the population is a parameter, and so is the variability among the unit values.

The objective of sample surveys is usually to estimate some parameter or a function of some parameter or parameters. Often, but not always, we wish to estimate the population mean or total. The value of the parameter as estimated from a sample will hereafter be referred to as the sample estimate or simply the estimate.

Bias, Accuracy, and Precision

In seeking an estimate of some population trait, the sampler's fondest hope is that at a reasonable cost he will obtain an estimate that is accurate (i.e., close to the true value). Without any help from sampling theory he knows that if bias rears its insidious head, accuracy will flee the scene. And he has a suspicion that even though bias is eliminated, his sample estimate may still not be entirely precise. When only a part of the population is measured, some estimates may be high, some low, some fairly close, and unfortunately, some rather far from the true value.

Though most people have a general notion as to the meaning of bias, accuracy, and precision, it might be well at this stage to state the statistical interpretation of these terms.

Bias.—Bias is a systematic distortion. It may be due to some flaw in measurement, to the method of selecting the sample, or to the technique of estimating the parameter. If, for example, seedling heights are measured with a ruler from which the first half-inch has been removed, all measurements will be one-half inch too large and the estimate of mean seedling height will be biased. In studies involving plant counts, some observers will nearly always include a plant that is on the plot boundary; others will 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. A table made up from tall timber will give biased results when used without adjustment on short-bodied trees. Similarly, if the cruiser consistently estimates merchantable height above or below the specifications of the table, volume so estimated will be biased. The only practical way to minimize measurement bias is by continual check of instrumentation, and meticulous training and care in the use of instruments.

Bias due to method of sampling may arise when certain units are given a greater or lesser representation in the sample than in the population. As an elementary example, assume that we are estimating the survival of 10,000 trees planted in 100 rows of 100 trees each. If the sample were selected only from the interior 98 x 98 block of trees in the interest of obtaining a "more representative" picture of survival, bias would occur simply because the border trees had no opportunity to appear in the sample.

The technique of estimating the parameter after the sample has been taken is also a possible source of bias. If, for example, the survival on a planting job is estimated by taking a simple arithmetic average of the survival estimates from two fields, the resulting average may be seriously biased if one field is 500 acres and the other 10 acres in size. A better overall estimate would be obtained by weighting the estimates for the two fields in proportion to the field sizes. Another example of this type of bias occurs in the common forestry practice of estimating average diameter 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.

Bias is seldom desirable, but it is not a cause for panic. It is something a sampler may have to live with. Its complete elimination may be costly in dollars, precision, or both. The important thing is to recognize the possible sources of bias and to weigh the effects against the cost of reducing or eliminating it. Some of the procedures discussed in this handbook are known to be slightly biased. They are used because the bias is often trivial and because they may be more precise than the unbiased procedures.

Precision and accuracy.—A badly biased estimate may be precise but it can never be accurate. Those who find this hard to swallow may be thinking of precision as being synonymous with accuracy. Statisticians being what they are, it will do little good to point out that several lexicographers seem to think the same way. Among statisticians *accuracy* refers to the success of estimating the true value of a quantity; *precision* refers to the clustering of sample values about their own average, which, if biased, cannot be the true value. Accuracy, or closeness to the true value, may be absent because of bias, lack of precision, or both.

A target shooter who puts all of his shots in a quarter-inch circle in the 10-ring might be considered accurate; his friend who puts all of his shots in a quarter-inch circle at 12 o'clock in the 6-ring would be considered equally precise but nowhere near as accurate. An example for foresters might be a series of careful measurements made of a single tree with a vernier caliper, one arm of which is not at right angles to the graduated beam. Because the measurements have been carefully made they should not vary a great deal but should cluster closely about their mean value: they will be precise. However, as 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 is properly adjusted but is used carelessly the measurements may be unbiased but they will be neither accurate nor precise.

Variables, Continuous and Discrete

Variation is one of the facts of life. It is difficult to say whether this is good or bad, but we can say that without it there would be no sampling problems (or statisticians). How to cope with some of the sampling problems created by natural variation is the subject of this handbook.

To understand statisticians it is helpful to know their language, and in this language the term variable plays an active part. A characteristic that may vary from unit to unit is called a variable. In a population of trees, tree height is a variable, so are tree diameter, number of cones, cubic volume, and form class. As some trees may be loblolly pine, some slash pine, and some dawn redwoods, species is also a variable. Presence or absence of insects, the color of the foliage, and the fact that the tree is alive or dead are variables also.

A variable that is characterized by being related to some numerical scale of measurement, any interval of which may, if desired, be subdivided into an infinite number of values, is said to be continuous. Length, height, weight, temperature, and volume are examples of variables that can usually be labeled continuous. 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. In the first of these the individual is classified as having or not having some attribute; or, more commonly, a group of individuals is described by the proportion or percentage having a particular attribute. Some familiar examples are the proportion of slash pine seedlings infected by rust, the percentage of stocked milacre quadrats, and the survival percentage of planted seedlings. In the second form, the individual is described by a count that cannot be expressed as a proportion. Number of seedlings on a milacre, number of weevils in a cone, number of sprouts on a stump, and number of female flowers on a tree are common examples.

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 described in this handbook were developed primarily for use with continuous variables. The procedures that have been devised for discrete variables are generally more complex. By increasing the number of values that a discrete variable can assume, however, it is often possible to handle such data by the continuous-variable methods. Thus, germination percentages based on 200 or more seeds per dish can usually be treated by the same procedures that would be used for measurement data. The section that begins on page 61 describes simple random sampling with classification data and gives some illustrations of how the sampling procedures for continuous data may be used for classification and count data.

Distribution Functions

A distribution function shows, for a population, the relative frequency with which different values of a variable occur. Knowing the distribution function, we can say what proportion of the individuals are within certain size limits.

Each population has its own distinct distribution function. There are, however, certain general types of function that occur quite frequently. The most common are the normal, binomial, and Poisson. The bell-shaped normal distribution, familiar to most foresters, is often encountered in dealing with continuous variables. The binomial is associated with data where a fixed number of individuals are observed on each unit and the unit is characterized by the number of individuals having some particular attribute. The Poisson distribution may arise where individual units are characterized by a count having no fixed upper limit, particularly if zero or very low counts tend to predominate.

The form of the distribution function dictates the appropriate statistical treatment of a set of data. The exact form of the distribution will seldom be known, but 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 which a variable follows, the means of large samples tend to follow a distribution that approaches the normal and may be treated by normal distribution methods.

TOOLS OF THE TRADE

Subscripts, Summations, and Brackets

In describing the various sampling methods, frequent use will be made of subscripts, brackets, and summation symbols. Some beginning samplers will be unhappy about this; others will be downright mad. The purpose though, is not to impress or confuse the reader. These devices are, like the more familiar notations of $+$, $-$, and $=$, merely a concise way of expressing ideas that would be ponderous if put into conventional language. And like the common algebraic symbols, using and understanding them is just a matter of practice.

Subscripts.—The appearance of an x_i, z_{jk}, or y_{ilmn} brings a frown of annoyance and confusion to the face of many a forester. Yet interpreting this notation is quite simple. In x_i, the subscript i means that x can take on different forms or values. Putting in a particular value of i tells which form or value of x we are concerned with. The i might imply a particular characteristic of an individual. The term x_1 might be the height of the individual, x_2 might be his weight, x_3 his 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 the height of the second, x_3 the height of the third individual, 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 with two subscripts. In x_{ik} the i might identify the characteristic (for height, $i = 1$; for weight, $i = 2$; and for age, $i = 3$). The k could be used to designate which 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 process 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 could write

$$(x_1 + x_2 + x_3 + x_4 + x_5 + x_6)$$

A slightly shorter way of saying the same thing is

$$(x_1 + x_2 + \ldots + x_6)$$

The three dots (...) indicate that we continue to do the same thing for all the values from x_3 through x_5 as we have already done to x_1 and x_2.

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 Σ, which is the Greek letter sigma, indicates that a summation should be performed. The x tells what is to be summed and the letter above and below Σ 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

$$\sum_{i} x_i, \ \sum x_i, \text{ or sometimes, } \sum x$$

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 would tell us to add up all the values of x_{ij} having j from 1 to 4 *and* i from 1 to 3. Written the long way, this means

$$(x_{1,1} + x_{1,2} + x_{1,3} + x_{1.4} + x_{2,1} + x_{2.2}$$
$$+ x_{2,3} + x_{2,4} + x_{3,1} + x_{3,2} + x_{3,3} + x_{3,4})$$

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

$$\sum_i \sum_j x_{ij}, \ \sum_{i,j} x_{ij}, \ \sum \sum x_{ij}, \ \sum x_{ij}, \text{ or maybe even } \sum x$$

If a numerical value is substituted for one of the letters in the subscript, the summation is to be performed by letting the letter subscript vary but holding the other subscript at the specified value. As an example,

$$\sum_{j=1}^{4} x_{3j} = (x_{3,1} + x_{3,2} + x_{3,3} + x_{3,4})$$

and,

$$\sum_{i=1}^{5} x_{i2} = (x_{1,2} + x_{2,2} + x_{3,2} + x_{4,2} + x_{5,2})$$

Bracketing.—When other operations are to be performed along with the addition, some form of 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

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

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

The expression

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

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

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

says 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

$$\sum_{i=1}^{3}\left(\sum_{j=1}^{4} x_{ij}\right)^2 = (x_{1,1} + x_{1,2} + x_{1,3} + x_{1,4})^2 + (x_{2,1} + x_{2,2} + x_{2,3} + x_{2,4})^2 + (x_{3,1} + x_{3,2} + x_{3,3} + x_{3,4})^2$$

The expression

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

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

$$\left(\sum_{i=1}^{3}\sum_{j=1}^{4} x_{ij}\right)^2 = (x_{1,1} + x_{1,2} + x_{1,3} + x_{1,4} + x_{2,1} + x_{2,2} + x_{2,3} + x_{2,4} + x_{3,1} + x_{3,2} + x_{3,3} + x_{3,4})^2$$

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,

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

N.B.: It is easily seen but often forgotten that

$$\sum_i x_i^2 \text{ is not usually equal to } \left(\sum_i x_i\right)^2$$

Similarly,

$$\sum_i x_i y_i \text{ is not usually equal to } \left(\sum_i x_i\right)\left(\sum_i y_i\right)$$

Some practice.—If you feel uncomfortable in the presence of this symbology, try the worked examples on page 79.

Variance

In a stand of trees, the diameters will usually show some variation. Some will be larger than the mean diameter, some smaller, and some fairly close to the mean. Clearly, it would be informative to know something about this variation. It is not hard to see that more observations would be needed to get a good estimate of the mean diameter in a stand where diameters vary from 2 to 30 inches than where the range is from 10 to 12 inches. The measure of variation most commonly used by statisticians is the *variance.*

The variance of individuals in a population is a measure of the dispersion of individual unit values about their mean. A large variance indicates wide dispersion, a small variance indicates little dispersion. The variance of individuals is a population characteristic (a parameter). Very rarely will we know the population variance. Usually it must be estimated from the sample data.

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

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

Where: s^2 = Sample estimate of the population variance.
y_i = The value of the i^{th} unit in the sample.
$\bar{y}$ = The arithmetic mean of the sample, i.e.,

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

n = The number of units observed in the sample.

Though it may not appear so, computation of the sample variance is simplified by rewriting the above equation as

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

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

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

The *standard deviation,* a term familiar to the survivors of most forest mensuration courses, is merely the square root of the variance. It is symbolized by s, and in the above example would be estimated as $s = \sqrt{7} = 2.6458$.

Standard Errors and Confidence Limits

Like the individual units in a population, sample estimates are subject to variation. The mean diameter of a stand as estimated from a sample of 3 trees will seldom be the same as the estimate that would have been obtained from other samples of 3 trees. One estimate might be close to the mean but a little high. Another might be quite a bit high, and the next might be below the mean. The estimates vary because different individual units are observed in the different samples.

Obviously, it would be 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 inches would inspire more confidence than one that might range from 6 to 18 inches.

The previous section discussed the variance and the standard deviation $\left(\text{standard deviation} = \sqrt{\text{variance}}\right)$ as 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 of the estimate and the standard error of estimate $\left(\text{standard error of estimate} = \sqrt{\text{variance of estimate}}\right)$. 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 among individual units. In fact, if several estimates were obtained by repeated sampling of a population, the variance and standard error of these estimates could be computed from the equations given in the previous section for the variance and standard deviation of individuals. But repeated sampling is unnecessary; the variance and standard error can be obtained from a single set of sample units. Variability of an estimate depends on the sampling method, the sample size, and the 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 in this handbook, the procedure for computing the standard error of estimate will be given.

Computation of a standard error is often regarded as an unnecessary frill by some self-styled practical foresters. The fact is, however, that a sample estimate is almost worthless without some indication of its reliability.

Given the standard error, it is possible to establish 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 inches with a standard error of 1.5 inches, we can say that the true mean is somewhere within the limits 14.5 to 17.5 inches. In making such statements we will, over the long run, be right an average of two times out of three. One time out of three we will, because of natural sampling variation, be wrong. 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 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 for a few of the sampling methods.

Expanded Variances and Standard Errors

Very often an estimate will be multiplied by a constant to put it in a more meaningful form. For example, if a survey has been made using one-fifth acre plots and the mean volume per plot computed, this estimate would be multiplied by 5 in order to put the estimated mean on a per acre basis. Or, for a tract of 800 acres the mean volume per fifth-acre plot would be multiplied by 4,000 (the number of one-fifth acres in the tract) in order to estimate the total volume.

Since expanding a variable 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 x has variance s^2 and this variable is multiplied by a constant (say k), the product (kx) will have a variance of k^2s^2.

Suppose the estimated mean volume per one-fifth acre plot is 1,400 board feet with a variance of 2,500 board feet (giving a standard error of $\sqrt{2{,}500} = 50$ board feet). The mean volume per acre is

$$\text{Mean volume per acre} = 5(1{,}400) = 7{,}000 \text{ board feet}$$

and the variance of this estimate is

$$\text{Variance of mean volume per acre} = (5^2)(2{,}500) = 62{,}500$$

The standard error of the mean volume per acre would be

$$\sqrt{\text{Variance of mean volume per acre}} = 250 \text{ board feet}$$

Note that if the standard deviation (or standard error) of x is s, then the standard deviation (or standard error) of kx is merely ks. So, in the above case, since the standard error of the estimated mean volume per fifth-acre plot is 50, the standard error of the mean volume per acre is $(5)(50) = 250$.

This is a simple but very important rule and anyone who will be dealing with sample estimates should master it.

Variables may also be expanded by the addition of a constant. Expansion of this type does not affect variability and requires no adjustment of the variance or standard errors. Thus if

$$z = x + k$$

where x is a variable and k a constant, then

$$s_z^2 = s_x^2$$

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 values are the same as for the uncoded values. Given the three observations 127, 104, and 114 we could, for ease of computation, code these values by subtracting 100 from each, to make 27, 4, and 14. The variance of the coded values is

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

which is the same as the variance of the original values

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

Coefficient of Variation

The coefficient of variation (C) is the ratio of the standard deviation to the mean. For a sample with a mean[1] of $\bar{x} = 10$ and a standard deviation of $s = 4$ we would estimate the coefficient of variation as

$$C = \frac{s}{\bar{x}} = \frac{4}{10} = 0.4 \text{ or } 40 \text{ percent}$$

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 foresters. The coefficient of variation puts the expression of variability on a relative basis. The population of tree heights might have a standard deviation of 4.4 feet while the population of foresters might have a standard deviation of 0.649 foot. In absolute units, the trees are more variable than the foresters. But, if the mean tree height is 40 feet and the mean height of the foresters is 5.9 feet, the two populations would have the same relative variability. They would both have a coefficient of variation of $C = 0.11$.

Variance also depends on the measurement units used. The standard deviation of foresters' heights was 0.649 foot. Had the heights been measured in inches, the standard deviation would have been 12 times as large (If $z = 12x \quad s_z = 12s_x$) or 7.788 inches. But the coefficient of variation would be the same regardless of the unit of measure. In either case, we would have

$$C = \frac{s}{\bar{x}} = \frac{0.649 \text{ foot}}{5.9 \text{ feet}} = \frac{7.788 \text{ inches}}{70.8 \text{ inches}} = 0.11 \text{ or } 11 \text{ percent}$$

[1] The sample mean of a variable x is frequently symbolized by $\bar{x}$.

In addition to putting variabilities 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 is approximately 100 percent. Because it is often possible to guess at the size of the population mean, we can readily estimate the standard deviation. Such information is useful in planning a sample survey.

Covariance

In some sampling methods measurements are made on two or more characteristics for each sample unit. In measuring forage production, for example, we might get the green weight of the grass clipped to a height of 1 inch from a circular plot 1 foot in diameter. Later we might get the ovendry weight of the same sample.

Covariance is a measure of how two variables vary in relationship to each other (covariability). 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 (s_{xy}) of x and y is

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

Computation of the sample covariance is simplified by rewriting the formula

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

Suppose that a sample of $n = 6$ units has produced the following x and y values:

i	1	2	3	4	5	6	Totals
y_i	2	12	7	14	11	8	54
x_i	12	4	10	3	6	7	42

Then,

$$s_{xy} = \frac{(2)(12) + (12)(4) + \ldots + (8)(7) - \left(\frac{(54)(42)}{6}\right)}{(6-1)}$$

$$= \frac{306 - 378}{5} = -14.4$$

The negative value indicates that the larger values of y tend to be associated with the smaller values of x.

Correlation Coefficient

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 (r) of the correlation coefficient is

$$r = \frac{\text{Covariance of } x \text{ and } y}{\sqrt{(\text{Variance of } x)(\text{Variance of } y)}} = \frac{s_{xy}}{\sqrt{(s_x^2)(s_y^2)}}$$

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

$$r_{xy} = \frac{-14.4}{\sqrt{(12.0)(18.4)}} = \frac{-14.4}{14.86} = -0.969$$

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.

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.

It must also be remembered 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, the reliability of a correlation coefficient increases with the sample size. Most statistics books have tables that help in judging the reliability of a sample correlation coefficient.

Independence

When no relationship exists between two variables they are said to be independent; the value of one variable tells us absolutely 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 very weakly related and may be regarded as independent. 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 50 to 120 square feet per acre). 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. For discussion purposes, two situations may be recognized:

Two estimates have been made of the same parameter.
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 acre of a timber stand. 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 the plot data from both the first and second surveys. Since some of the same observations enter both estimates, the estimates would not be independent. 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 proportions of observations common to the two estimates.

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 set were selected 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 (covariance) between two sample estimates is essential in the evaluation of the sampling error for several types of surveys. For the sampling methods described in this handbook, the procedure for computing the covariance of two estimates will be given when needed.

Variances of Products, Ratios, and Sums

In a previous section, 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, if we wish to estimate the total volume of a stand, we would 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 $Ns_{\bar{y}}$.

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,

$$Q_1 = kzw$$

where: k = a constant and

z and w = variables having variances s_z^2 and s_w^2 and covariance s_{zw}

For large samples, the variance of Q_1 is estimated by

$$s_{Q_1}{}^2 = Q_1{}^2 \left(\frac{s_z{}^2}{z^2} + \frac{s_w{}^2}{w^2} + \frac{2s_{wz}}{zw} \right)$$

As an example of such estimates, consider a large 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 acre ($\bar{v}$) of forested land. To estimate the forested acreage, the total acreage (N) in the area is multiplied by the estimated proportion forested. This in turn is multiplied by the mean volume per forested acre to give the total volume. In formula form

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

Where: N = The total acreage of the area (a known constant).
$\bar{p}$ = The estimated proportion of the area that is forested.
$\bar{v}$ = The estimated mean volume per forested acre.

The variance of the estimated total volume would be

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

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 here 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

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

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

$$Q_2 = k\frac{z}{w}$$

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

$$s_{Q_2}{}^2 = Q_2{}^2 \left[\frac{s_z{}^2}{z^2} + \frac{s_w{}^2}{w^2} - \frac{2s_{zw}}{zw}\right]$$

This formula comes into use with the ratio-of-means estimator described in the section on regression estimators.

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

$$Q_3 = k_1x_1 + k_2x_2$$

where: k_1 and k_2 = constants
x_1 and x_2 = variables having variance $s_1{}^2$ and $s_2{}^2$ and covariance s_{12}

The variance of this estimate is

$$s_{Q_3}{}^2 = k_1{}^2s_1{}^2 + k_2{}^2s_2{}^2 + 2k_1k_2s_{12}$$

If we measure the volume of sawtimber (x) and the volume of poletimber (y) on the same plots (and in the same units of measure) and find the mean volumes to be $\bar{x}$ and $\bar{y}$, with variances $s_{\bar{x}}{}^2$ and $s_{\bar{y}}{}^2$ and covariance $s_{\bar{x}\bar{y}}$, then the mean total volume in pole-size and larger trees would be

$$\text{Mean total volume} = \bar{x} + \bar{y}$$

The variance of this estimate is

$$s^2 = s_{\bar{x}}{}^2 + s_{\bar{y}}{}^2 + 2s_{\bar{x}\bar{y}}$$

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.

Tract 1
Size = 3,200 acres
Estimated mean volume per acre = 4,800 board feet
Variance of the mean = 112,500 board feet

Tract 2
Size = 1,200 acres
Estimated mean volume per acre = 7,400 board feet
Variance of the mean = 124,000 board feet

In combining these two means to estimate the overall mean volume per acre 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 acre). Thus,

$$\begin{aligned}\bar{x} &= \frac{3200(4800) + 1200(7400)}{(3200 + 1200)} \\ &= \left(\frac{3200}{4400}\right)(4800) + \left(\frac{1200}{4400}\right)(7400) = 5509\end{aligned}$$

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

$$\begin{aligned}s_{\bar{x}}^2 &= \left(\frac{3200}{4400}\right)^2 (112{,}500) + \left(\frac{1200}{4400}\right)^2 (124{,}000) \\ &= \frac{(3200)^2(112{,}500) + (1200)^2(124{,}000)}{(4400)^2} \\ &= 68{,}727.\end{aligned}$$

The general rule for the variance of a sum is if

$$Q = k_1x_1 + k_2x_2 + \ldots + k_nx_n$$

where: k_i = constants
x_i = variables with variances s_i^2 and covariance s_{ij},
then

$$\begin{aligned}s_Q^2 = k_1^2s_1^2 + k_2^2s_2^2 + \ldots + k_n^2s_n^2 + 2k_1k_2s_{12} \\ + 2k_1k_3s_{13} + \ldots + 2k_{n\text{-}1}\ k_ns_{(n\text{-}1)n}\end{aligned}$$

Transformation of Variables

Many of the procedures described in this handbook 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 (transformed).

One of the common assumptions is that variability is independent of the mean. Some variables (e.g., those that follow a binomial or Poisson distribution) tend to have a variance that is in some way related to the mean—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, these 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. The arcsin transformation is illustrated on page 66.

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. A variable may also be transformed to convert its distribution to the normal on which many of the simpler statistical methods are based.

The amateur sampler will do well to seek expert advice when transformations are being considered.

Finally, it should be noted that transformation is not synonymous with coding, which is done to simplify computation. Nor is it a form of mathematical hocus-pocus aimed at obtaining answers that are in agreement with preconceived notions.

SAMPLING METHODS FOR CONTINUOUS VARIABLES

Simple Random Sampling

All of the sampling methods to be described in this handbook have their roots in simple random sampling. Because it is basic, the method will be discussed in greater detail than any of the other procedures.

The fundamental idea in simple random sampling is that, in choosing a sample of n units, every possible combination of n units should have an equal chance of being selected. This is not the same as requiring that every unit in the population have an equal chance of being selected. The latter requirement is met by many forms of restricted randomization and even by some systematic selection methods.

Giving every possible combination of n units an equal chance of appearing in a sample of size n may be difficult to visualize but is easily accomplished. It is only necessary to be sure that at any stage of the sampling the selection of a particular unit is in no way influenced by the other units that have been selected. To state it in another way, the selection of any given unit should be completely independent of the selection of all other units. One way to do this is to assign every unit in the population a number and then draw n numbers from a table of random digits (table 1, p. 82). Or, the numbers can be written on some equal-sized disks or slips of paper which are placed in a bowl, thoroughly mixed, and then drawn one at a time. For units such as individual tree seeds, the units themselves may be drawn at random.

The units may be selected with or without replacement. If selection is with replacement, each unit is allowed to appear in the sample as often as it is selected. In sampling without replacement, a particular unit is allowed to appear in the sample only once. Most forest sampling is without replacement. As will be shown later, the procedure for computing standard errors depends on whether sampling was with or without replacement.

Sample selection.—The selection method and computations may be illustrated by the sampling of a 250-acre plantation. The objective of the survey was to estimate the mean cordwood volume per acre in trees more than 5 inches d.b.h. outside bark. The population and sample units were defined to be square quarter-acre plots with the unit value being the plot volume. The sample was to consist of 25 units selected at random and without replacement.

The quarter-acre units were plotted on a map of the plantation and assigned numbers from 1 to 1,000. From a table of random digits, 25 three-digit numbers were selected to identify the units to be included in the sample (the number 000 was associated with the plot numbered 1,000). No unit was counted in the sample more than once. Units drawn a second time were rejected and an alternative unit was randomly selected.

The cordwood volumes measured on the 25 units were as follows:

7	10	7	4	7
8	8	8	7	5
2	6	9	7	8
6	7	11	8	8
7	3	8	7	7
			Total =	175

Estimates.—If the cordwood volume on the i^{th} sampling unit is designated y_i, the estimated mean volume ($\bar{y}$) per sampling unit is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{7 + 8 + 2 + \ldots + 7}{25} = \frac{175}{25}$$

$$= 7 \text{ cords per quarter-acre plot.}$$

The mean volume per acre would, of course, be 4 times the mean volume per quarter-acre plot, or 28 cords.

As there is a total of $N = 1{,}000$ quarter-acre units in the 250-acre plantation, the estimated total volume ($\hat{Y}$) in the plantation would be

$$\hat{Y} = N\bar{y} = (1{,}000)\,(7) = 7{,}000 \text{ cords.}$$

Alternatively,

$$\hat{Y} = (28 \text{ cords per acre})\,(250 \text{ acres}) = 7{,}000 \text{ cords.}$$

Standard errors.—A first step in computing the standard error of estimate is to make an estimate ($s_y{}^2$) of the variance of individual values of y.

$$s_y{}^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{(n-1)}$$

In this example,

$$s_y{}^2 = \frac{(7^2 + 8^2 + \ldots + 7^2) - \frac{(175)^2}{25}}{(25-1)}$$

$$= \frac{1{,}317 - 1{,}225}{24} = 3.8333 \text{ cords}$$

When sampling is without replacement the standard error of the mean ($s_{\bar{y}}$) for a simple random sample is

$$s_{\bar{y}} = \sqrt{\frac{s_y{}^2}{n}\left(1 - \frac{n}{N}\right)}$$

where: N = total number of sample units in the entire population,
n = number of units in the sample.

For the plantation survey,

$$s_{\bar{y}} = \sqrt{\frac{3.8333}{25}\left(1 - \frac{25}{1{,}000}\right)} = \sqrt{(.1533)\,(.975)}$$

$$= 0.387 \text{ cord}$$

This is the standard error for the mean per quarter-acre plot. By the rules for the expansion of variances and standard errors, the standard error for the mean volume per acre will be (4) (0.387) = 1.548 cords.

Similarly, the standard error for the estimated total volume ($s_{\hat{Y}}$) will be

$$s_{\hat{Y}} = Ns_{\bar{y}} = (1{,}000)\,(.387) = 387 \text{ cords.}$$

Sampling with replacement.—In the formula for the standard error of the mean, the term $(1 - \frac{n}{N})$ is known as the finite population correction or fpc. It is used when units are selected without replacement. If units are selected with replacement, the fpc is omitted and the formula for the standard error of the mean becomes

$$s_{\bar{y}} = \sqrt{\frac{s_y^2}{n}}$$

Even when sampling is without replacement the sampling fraction (n/N) may be extremely small, making the fpc very close to unity. If n/N is less than 0.05, the fpc is commonly ignored and the standard error computed from the shortened formula.

Confidence limits for large samples.—By itself, the estimated mean of 28 cords per acre does not tell us very much. Had the sample consisted of only 2 observations we might conceivably have drawn the quarter-acre plots having only 2 and 3 cords, and the estimated mean would be 10 cords per acre. Or if we had selected the plots with 10 and 11 cords, the mean would be 42 cords per acre.

To make an estimate meaningful it is necessary to compute confidence limits that indicate the range within which we might expect (with some specified degree of confidence) to find the parameter. As was discussed in the chapter on standard errors, the 95-percent confidence limits for large samples are given by

$$\text{Estimate} \pm 2 \text{ (Standard Error of Estimate)}$$

Thus the mean volume per acre (28 cords) that had a standard error of 1.548 cords would have confidence limits of

$$28 \pm 2(1.548) = 24.90 \text{ to } 31.10 \text{ cords per acre.}$$

And the total volume of 7,000 cords that had a standard error of 387 cords would have 95-percent confidence limits of

$$7{,}000 \pm 2(387) = 6{,}226 \text{ to } 7{,}774 \text{ cords.}$$

Unless a 1-in-20 chance has occurred in sampling, the population mean volume per acre is somewhere between 24.9 and 31.1 cords, and the true total volume is between 6,226 and 7,774 cords.

Because of sampling variation, the 95-percent confidence limits will, on the average, fail to include the parameter in 1 case out of 20. It must be emphasized, however, that *these limits and the confidence statement take account of sampling variation only.* They assume that the plot values are without measurement error and that the sampling and estimating procedures are unbiased and free of computational mistakes. If these basic assumptions are not valid, the estimates and confidence statements may be nothing more than a statistical hoax.

Confidence limits for small samples.—Ordinarily, large-sample confidence limits are not appropriate for samples of less than 30 observations. For smaller samples the proper procedure depends on the distribution of the unit values in the parent population, a subject that is beyond the scope of this handbook. Fortunately, many forest measurements follow the bell-shaped normal distribution, or a distribution that can be made nearly normal by transformation of the variable.

For samples of any size from *normally distributed populations,* Student's t value can be used to compute confidence limits. The general formula is

$$\text{Estimate} \pm (t)\ (\text{Standard Error of Estimate}).$$

The values of t have been tabulated (table 2, page 86). The particular value of t to be used depends on the degree of confidence desired and on the size of the sample. For 95-percent confidence limits, the t values are taken from the column for a probability of .05. For 99-percent confidence limits, the t value would come from the .01 probability column. Within the specified columns, the appropriate t for a simple random sample of n observations is found in the row for $(n - 1)$ df's (degrees of freedom[2]). For a simple random sample of 25 observations the t value for computing the 95-percent confidence limits will be found in the .05 column and the 24 df row. This value is 2.064. Thus, for the plantation survey that showed a mean per-acre volume of 28 cords and a standard error of the mean of 1.548 cords, the small-sample 95-percent confidence limits would be

$$28 \pm (2.064)\ (1.548) = 24.80 \text{ to } 31.20 \text{ cords}$$

The same t value is used for computing the 95-percent confidence limits on the total volume. As the estimated total was 7,000 cords with a standard error of 387 cords, the 95-percent confidence limits are

$$7{,}000 \pm (2.064)\ (387) = 6{,}201 \text{ to } 7{,}799 \text{ cords.}$$

Size of sample.—In the example illustrating simple random sampling, 25 units were selected. But why 25? Why not 100? Or 10? All too often the number depends on the sampler's view of what looks about right. But there is a somewhat more objective solution. That is to take only the number of observations needed to give the desired precision.

In planning the plantation survey, we could have stated that unless a 1-in-20 chance occurs we would like our sample estimate of the mean to be within $\pm\ E$ cords of the population mean. As the small-sample confidence limits are computed as $\bar{y} \pm t(s_{\bar{y}})$, this is equivalent to saying that we want

$$t(s_{\bar{y}}) = E$$

For a simple random sample

$$s_{\bar{y}} = \sqrt{\frac{s_y^2}{n}\left(1 - \frac{n}{N}\right)}$$

[2] In this handbook the expression "degrees of freedom" refers to a parameter in the distribution of Student's t. When a tabular value of t is required, the number of degrees of freedom (df's) must be specified. The expression is not easily explained in nonstatistical language. One definition is that the df's are equal to the number of observations in a sample minus the number of independently estimated parameters used in calculating the sample variance. Thus, in a simple random sample of n observations the only estimated parameter needed in calculating the sample variance is the mean (x), and so the df's would be $(n - 1)$.

Substituting for $s_{\bar{y}}$ in the first equation we get

$$(t)\sqrt{\frac{s_y^2}{n}\left(1-\frac{n}{N}\right)}=E$$

Rewritten in terms of the sample size (n) this becomes

$$n=\cfrac{1}{\cfrac{E^2}{t^2s_y^2}+\cfrac{1}{N}}$$

To solve this relationship for n, we must have some estimate (s_y^2) of the population variance. Sometimes the information is available from previous surveys. In the illustration, we found $s_y^2 = 3.83$, a value which might be taken as representative of the variation among quarter-acre plots in this or similar populations. In the absence of this information, a small preliminary survey might be made in order to obtain an estimate of the variance. When, as often happens, neither of these solutions is feasible, a very crude estimate can be made from the relationship

$$s_y^2=\left(\frac{R}{4}\right)^2$$

where: R = estimated range from the smallest to the largest unit value likely to be encountered in sampling.

For the plantation survey we might estimate the smallest y-value on quarter-acre plots to be 1 cord and the largest to be 10 cords. As the range is 9, the estimated variance would be

$$s_y^2=\left(\frac{9}{4}\right)^2=5.06$$

This approximation procedure should be used only when no other estimate of the variance is available.

Having specified a value of E and obtained an estimate of the variance, the last piece of information we need is the value of t. Here we hit somewhat of a snag. To use t we must know the number of degrees of freedom. But, the number of df's must be $(n - 1)$ and n is not known and cannot be determined without knowing t.

An iterative solution will give us what we need, and it is not as difficult as it sounds. The procedure is to guess at a value of n, use the guessed value to get the degrees of freedom for t, and then substitute the appropriate t value in the sample-size formula to solve for a first approximation of n. Selecting a new n somewhere between the guessed value and the first approximation, but closer to the latter, we compute a second approximation. The process is repeated until successive values of n are the same or only slightly different. Three trials usually suffice.

To illustrate the process, suppose that in planning the plantation survey we had specified that, barring a 1-in-100 chance, we would like the estimate to be within 3.0 cords of the true mean volume per acre. This is equivalent to $E = 0.75$ cord per quarter-acre. From previous experience, we estimate the population variance among quarter-acre plots to be $s_y^2 = 4$, and we know that there is a total of $N = 1{,}000$ units in the population. To solve for n, this information is substituted in the sample-size formula given on page 25.

$$n = \frac{1}{\frac{(0.75)^2}{(t^2)(4)} + \frac{1}{1{,}000}}$$

We will have to use the t value for the .01 probability level, but we do not know how many degrees of freedom t will have without knowing n. As a first guess, we can try $n = 61$; then the value of t with 60 degrees of freedom at the .01 probability level is $t = 2.66$. Thus, the first approximation will be

$$n_1 = \frac{1}{\frac{(0.75^2)}{(2.66^2)(4)} + \frac{1}{1{,}000}} = \frac{1}{\frac{.5625}{(7.0756)(4)} + \frac{1}{1.000}}$$
$$= 47.9$$

A second guessed value for n would be somewhere between 61 and 48, but closer to the computed value. We might test $n = 51$, for which the value of t (50 df's) at the .01 level is about 2.68, whence

$$n_2 = \frac{1}{\frac{.5625}{(7.1824)(4)} + \frac{1}{1{,}000}}$$
$$= 48.6$$

The desired value is somewhere between 51 and 48.6 but much closer to the latter. Because the estimated sample size is, at best, only a good approximation, it is rather futile to strain on the computation of n. In this case we would probably settle on $n = 50$, a value that could have been easily guessed after the first approximation was computed.

If the sampling fraction $\frac{n}{N}$ is likely to be small (say, less than 0.05), the finite-population correction $\left(1 - \frac{n}{N}\right)$ may be ignored in the estimation of sample size and the formula simplified to

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

This formula is also appropriate in sampling with replacement. In the previous example the simplified formula gives an estimated sample size of $n = 51$.

The short formula is frequently used to get a first approximation of n. Then, if the sample size indicated by the short formula is a considerable proportion (say over 10 percent) of the number of units in the population and sampling will be without replacement, the estimated sample size is recomputed with the long formula.

Effect of plot size on variance.—In estimating sample size, the effect of plot size and the scale of the unit values on variance must be kept in mind. In the plantation survey a plot size of one-quarter acre was selected and the variance among plot volumes was estimated to be $s^2 = 4$. This is the variance among volumes per quarter-acre. Because the desired precision was expressed on a per-acre basis it was necessary to modify either the precision specification or s^2 to get them on the same scale. In the example, s^2 was used without change and the desired precision was divided by 4 to put it on a quarter-acre basis. The same result could have been obtained by leaving the specified precision unchanged and putting the variance on a per-acre basis. Since the quarter-acre volumes would be multiplied by 4 to put them on a per-acre basis, the variance of quarter-acre volumes should be multiplied by 16. (Remember: If x is a variable with variance s^2, then the variance of a variable $z = kx$ is k^2s^2).

Plot size has an additional effect on variance. At the same scale of measurement, small plots will almost always be more variable than large ones. The variance in volume per acre on quarter-acre plots would be somewhat larger than the variance in volume per acre on half-acre plots, but slightly smaller than the variance in volume per acre of fifth-acre plots. Unfortunately, the relation of plot size to variance changes from one population to another. Large plots tend to have a smaller variance because they average out the effect of clumping and holes. In very uniform populations, changes in plot size have little effect on variance. In nonuniform populations the relationship of plot size to variance will depend on how the sizes of clumps and holes compare to the plot sizes. Experience is the best guide as to the effect of changing plot size on variance. Where neither experience nor advice is available, a very rough approximation can be obtained by the rule:

> If plots of size P_1 have a variance s_1^2 then, on the same scale of measurement, plots of size P_2 will have a variance roughly equal to

$$s_2^2 = s_1^2 \sqrt{P_1/P_2}$$

Thus, if the variance in cordwood volume *per acre* on quarter-acre plots is $s_1^2 = 61$, the variance in cordwood volume *per acre* on tenth-acre plots will be roughly

$$61 \sqrt{0.25/0.10} = 96$$

The same results will be obtained without worry about the scale of measurement if the squared coefficients of variation (C^2) are used in place of the variances. The formula would then be

$$C_2^2 = C_1^2 \sqrt{P_1/P_2}$$

Practice problem.—A survey is to be made to estimate the mean board-foot volume per acre in a 200-acre tract. Barring a 1-in-20 chance, we would like the estimate to be within 500 board feet of the population mean. Sample plots will be one-fifth acre. A survey in a similar tract showed the standard deviation among quarter-acre plot volumes to be 520 board feet. What size sample will be needed?

Problem Solution:

The variance among quarter-acre plot volumes is $520^2 = 270{,}400$. For quarter-acre volumes expressed on a per-acre basis the variance would be

$$s_1^2 = (4^2)\,(270{,}400) = 4{,}326{,}400$$

The estimated variance among fifth-acre plot volumes expressed on a per-acre basis would then be

$$\begin{aligned} s_2^2 &= s_1^2\sqrt{\frac{P_1}{P_2}} = 4{,}326{,}400\sqrt{\frac{0.25}{0.20}} \\ &= (4{,}326{,}400)\,(1.118) \\ &= 4{,}836{,}915 \end{aligned}$$

The population size is $N = 1{,}000$ fifth-acre plots.

If as a first guess $n = 61$, the t value at the .05 level with 60 degrees of freedom is 2.00. The first computed approximation of n is

$$n_1 = \frac{1}{\dfrac{(500)^2}{(4)\,(4{,}836{,}915)} + \dfrac{1}{1{,}000}} = 71.8$$

The correct solution is between 61 and 71.8 but much closer to the computed value. Repeated trials will give values between 71.0 and 71.8. The sample size (n) must be an integral value and, because 71 is too small, a sample of $n = 72$ observations would be required for the desired precision.

Stratified Random Sampling

Often we have knowledge of a population which can be used to increase the precision or usefulness of our sample. Stratified random sampling is a method that takes advantage of certain types of information about the population.

In stratified random sampling, the units of the population are grouped together on the basis of similarity of some characteristic. Each group or stratum is then sampled and the group estimates are combined to give a population estimate.

In sampling a forest, we might set up strata corresponding to the major timber types, make separate sample estimates for each type, and then combine the type data to give an estimate for the entire population. If the variation among units within types is less than the variation among units that are not in the same type, the population estimate will be more precise than if sampling had been at random over the entire population.

The sampling and computational procedures can be illustrated with data from a cruise made to estimate the mean cubic-foot volume per acre on an 800-acre forest. On aerial photographs the tract was divided into three strata corresponding to three major forest types; pine, bottom-land hardwoods, and upland hardwoods. The boundaries and total acreage of each type were known. Ten one-acre plots were selected at random and without replacement in each stratum.

Stratum		*Observations*		
I. Pine	570	510	600	
	640	590	780	Total = 6,100
	480	670	700	
	560			
II. Bottom-land hardwoods	520	630	810	
	710	760	580	Total = 7,370
	770	890	860	
	840			
III. Upland hardwoods	420	540	320	
	210	180	270	Total = 3,040
	290	260	200	
	350			

Estimates.—The first step in estimating the population mean per unit is to compute the sample mean ($\bar{y}_h$) for each stratum. The procedure is the same as for the mean of a simple random sample.

$\bar{y}_I = 6{,}100/10 = 610$ cubic feet per acre for the pine type
$\bar{y}_{II} = 7{,}370/10 = 737$ cubic feet per acre for bottom-land hardwoods
$\bar{y}_{III} = 3{,}040/10 = 304$ cubic feet per acre for upland hardwoods

The mean of a stratified sample ($\bar{y}_{st}$) is then computed by

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

Where: L = The number of strata.

N_h = The total size (number of units) of stratum h ($h=1, \ldots, L$).

N = The total number of units in all strata $\left(N = \sum_{h=1}^{L} N_h\right)$.

If the strata sizes are

I.	Pine	= 320 acres	$= N_I$
II.	Bottom-land hardwoods	= 140 acres	$= N_{II}$
III.	Upland hardwoods	= 340 acres	$= N_{III}$
	Total	= 800 acres	$= N$

Then the estimate of the population mean is

$$\bar{y}_{st} = \frac{(320)(610) + (140)(737) + (340)(304)}{800}$$
$$= 502.175 \text{ cubic feet per acre}$$

For the estimate of the population total ($\hat{Y}_{st}$), simply omit the divisor N.

$$\hat{Y}_{st} = \sum_{h=1}^{L} N_h \bar{y}_h = 320(610) + 140(737) + 340(304) = 401{,}740$$

Alternatively,

$$\hat{Y}_{st} = N\bar{y}_{st} = 800(502.175) = 401{,}740$$

Standard errors.—To determine standard errors, it is first necessary to obtain the estimated variance among individuals within each stratum (s_h^2). These variances are computed in the same manner as the variance of a simple random sample. Thus, the variance within Stratum I (Pine) is

$$s_I^2 = \frac{(570^2 + 640^2 + \ldots + 700^2) - \frac{(6100)^2}{10}}{(10 - 1)}$$
$$= \frac{3{,}794{,}000 - 3{,}721{,}000}{9}$$
$$= 8111.1111$$

Similarly,

$$s_{II}^2 = 15{,}556.6667$$
$$s_{III}^2 = 12{,}204.4444$$

From these values we find the standard error of the mean of a stratified random sample ($s_{\bar{y}_{st}}$) by the formula

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

Where: n_h = Number of units observed in stratum h.

This looks rather ferocious and does get to be a fair amount of work, but it is not too bad if taken step by step. For the timber cruising example we would have

$$\begin{aligned} s_{\bar{y}_{st}} &= \sqrt{\frac{1}{800^2}\left[\frac{(320)^2(8111.1111)}{10}\left(1-\frac{10}{320}\right) + \cdots + \frac{(340)^2(12{,}204.4444)}{10}\left(1-\frac{10}{340}\right)\right]} \\ &= \sqrt{383.920659} \\ &= 19.594 \end{aligned}$$

As a rough rule we can say that unless a 1-in-20 chance has occurred, the population mean is included in the range

$$\begin{aligned} \bar{y}_{st} \pm 2\ (s_{\bar{y}_{st}}) &= 502.175 \pm 2(19.594) \\ &= 463 \text{ to } 541 \end{aligned}$$

If sampling is with replacement or if the sampling fraction within a particular stratum (n_h/N_h) is small, we can omit the finite-population correction $\left(1 - \frac{n_h}{N_h}\right)$ for that particular stratum when calculating the standard error.

The population total being estimated by $\hat{Y}_{st} = N\bar{y}_{st}$, the standard error of $\hat{Y}_{st}$ is simply

$$s_{\hat{Y}_{st}} = Ns_{\bar{y}_{st}} = 800(19.594) = 15{,}675$$

Discussion.—Stratified random sampling offers two primary advantages over simple random sampling. First, it provides separate estimates of the mean and variance of each stratum. Second, for a given sampling intensity, it often gives more precise estimates of the population parameters than would a simple random sample of the same size. For this latter advantage, however, it is necessary that the strata be set up so that the variability among unit values within the strata is less than the variability among units that are not in the same stratum.

Some drawbacks are that each unit in the population must be assigned to one and only one stratum, that the size of each stratum must be known, and that a sample must be taken in each stratum. The most common barrier to the use of stratified random sampling is lack of knowledge of the strata sizes. If the sampling fractions are small in each stratum, it is not necessary to know the exact strata sizes; the population mean and its standard error can be computed from the relative sizes. If r_h = the relative size of stratum h, the estimated mean is

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} r_h \bar{y}_h}{\sum_{h=1}^{L} r_h}$$

The estimated standard error of the mean is

$$s_{\bar{y}_{st}} = \sqrt{\frac{\sum_{h=1}^{L}\left(\frac{r_h{}^2 s_h{}^2}{n_h}\right)}{\left(\sum_{h=1}^{L} r_h\right)^2}}$$

It is worth repeating that the sizes or relative sizes of the strata must be known in advance of sampling; the error formulae given above are not applicable if the observations from which the strata means are estimated are also used to estimate the strata sizes.

Sample Allocation in Stratified Random Sampling

Assuming we have decided on a total sample size of n observations, how do we know how many of these observations to make in each stratum? Two common solutions to this problem are known as proportional and optimum allocation.

Proportional allocation.—In this procedure the proportion of the sample that is selected in the h^{th} stratum is made equal to the proportion of all units in the population which fall in that stratum. If a stratum contains half of the units in the population, half of the sample observations would be made in that stratum. In equation form, if the total number of sample units is to be n, then for proportional allocation the number to be observed in stratum h is

$$n_h = \left(\frac{N_h}{N}\right) n$$

In the previous example, the 30 sample observations were divided equally among the strata. For proportional allocation we would have used

$$n_I = \left(\frac{N_I}{N}\right) n = \left(\frac{320}{800}\right) 30 = 12$$

$$n_{II} = \left(\frac{140}{800}\right) 30 = 5.25 \text{ or } 5$$

$$n_{III} = \left(\frac{340}{800}\right) 30 = 12.75 \text{ or } 13$$

Optimum allocation.—In optimum allocation the observations are allocated to the strata so as to give the smallest standard error possible with a total of n observations. For a sample of size n, the number of observations (n_h) to be made in stratum h under optimum allocation is

$$n_h = \left(\frac{N_h s_h}{\sum_{h=1}^{L} N_h s_h}\right) n$$

In terms of the previous example the value of $N_h s_h$ for each stratum is

$$\begin{aligned}
N_I s_I &= 320\sqrt{8111.1111} = 320\,(90.06) = 28{,}819.20 \\
N_{II} s_{II} &= 140\sqrt{15{,}556.6667} = 140\,(124.73) = 17{,}462.20 \\
N_{III} s_{III} &= 340\sqrt{12{,}204.4444} = 340\,(110.47) = 37{,}559.80 \\
&\text{Total} = \overline{83{,}841.20} = \sum_{h=I}^{III} N_h s_h
\end{aligned}$$

Applying these values in the formula, we would get

$$\begin{aligned}
n_I &= \left(\frac{28{,}819.20}{83{,}841.20}\right) 30 = 10.3 \text{ or } 10 \\
n_{II} &= \left(\frac{17{,}462.20}{83{,}841.20}\right) 30 = 6.2 \text{ or } 6 \\
n_{III} &= \left(\frac{37{,}559.80}{83{,}841.20}\right) 30 = 13.4 \text{ or } 14
\end{aligned}$$

Here optimum allocation is not much different from proportional allocation. Sometimes the difference is great.

Optimum Allocation With Varying Sampling Costs

Optimum allocation as just described assumes that the sampling cost per unit is the same in all strata. When sampling costs vary from one stratum to another, the allocation giving the most information per dollar is

$$n_h = \left(\frac{\dfrac{N_h s_h}{\sqrt{c_h}}}{\Sigma\left(\dfrac{N_h s_h}{\sqrt{c_h}}\right)} \right) n$$

Where: c_h = Cost per sampling unit in stratum h.

The best way to allocate a sample among the various strata depends on the primary objectives of the survey and our information about the population. One of the two forms of optimum allocation is preferable if the objective is to get the most precise estimate of the population mean for a given cost. If we want separate estimates for each stratum and the overall estimate is of secondary importance, we may want to sample heavily in the strata having high-value material. Then we would ignore both optimum and proportional allocation and place our observations so as to give the degree of precision desired for the particular strata.

We cannot, of course, use optimum allocation without having some idea about the variability within the various strata. The appropriate measure of variability within the stratum is the standard deviation (not the standard error), but we need not know the exact standard deviation (s_h) for each stratum. In place of actual s_h values, we can use relative values. In our example, if we had known that the standard deviations for the strata were about in the proportions $s_I : s_{II} : s_{III} = 9:12:11$, we could have used these values and obtained about the same allocation. Where optimum allocation is indicated but nothing is known about the strata standard deviations, proportional allocation is often very satisfactory.

Caution! In some situations the optimum allocation formula will indicate that the number of units (n_h) to be selected in a stratum is larger than the stratum (N_h) itself. The common procedure then is to sample all units in the stratum and to recompute the total sample size (n) needed to obtain the desired precision. The method of estimating n is discussed in the next section.

Sample Size in Stratified Random Sampling

In order to estimate the total size of sample (n) needed in a stratified random sample, the following pieces of information are required:

A statement of the desired size of the standard error of the mean. This will be symbolized by D.

A reasonably good estimate of the variance (s_h^2) or standard deviation (s_h) among individuals within each stratum.

The method of sample allocation. If the choice is optimum allocation with varying sampling costs, the sampling cost per unit for each stratum must also be known.

Given this hard-to-come-by information, we can estimate the size of sample (n) with these formulae:

For equal samples in each of the L strata,

$$n = \frac{L \sum_{h=1}^{L} N_h^2 s_h^2}{N^2 D^2 + \sum_{h=1}^{L} N_h s_h^2}$$

For proportional allocation,

$$n = \frac{N \sum_{h=1}^{L} N_h s_h^2}{N^2 D^2 + \sum_{h=1}^{L} N_h s_h^2}$$

For optimum allocation with equal sampling costs among strata,

$$n = \frac{\left(\sum_{h=1}^{L} N_h s_h \right)^2}{N^2 D^2 + \sum_{h=1}^{L} N_h s_h^2}$$

For optimum allocation with varying sampling costs among strata,

$$n = \frac{\left(\sum_{h=1}^{L} N_h s_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} \frac{N_h s_h}{\sqrt{c_h}}\right)}{N^2D^2 + \sum_{h=1}^{L} N_h s_h^2}$$

When the sampling fractions $\left(\frac{n_h}{N_h}\right)$ are likely to be very small for all strata or when sampling will be with replacement, the second term of the denominators of the above formulae $\left(\sum_{h=1}^{L} N_h s_h^2\right)$ may be omitted, leaving only N^2D^2.

If the optimum allocation formula indicates a sample (n_h) greater than the total number of units (N_h) in a particular stratum, n_h is usually made equal to N_h; i.e., all units in that particular stratum are observed. The previously estimated sample size (n) should then be dropped and the total sample size (n') and allocation for the remaining strata recomputed omitting the N_h and s_h values for the offending stratum but leaving N and D unchanged.

As an illustration, assume a population of 4 strata with sizes (N_h) and estimated variances s_h^2 as follows:

Stratum	N_h	s_h^2	s_h	$N_h s_h$	$N_h s_h^2$
1	200	400	20	4,000	80,000
2	100	900	30	3,000	90,000
3	400	400	20	8,000	160,000
4	20	19,600	140	2,800	392,000
	$N =$ 720			17,800	722,000

With optimum allocation (same sampling cost per unit in all strata), the number of observations to estimate the population mean with a standard error of $D = 1$ is

$$n = \frac{(17{,}800)^2}{(720^2)(1^2) + 722{,}000} = 255.4 \text{ or } 256$$

The allocation of these observations according to the optimum formula would be

$$n_1 = \left(\frac{4{,}000}{17{,}800}\right) 256 = 57.5 \text{ or } 58$$

$$n_2 = \left(\frac{3{,}000}{17{,}800}\right) 256 = 43.1 \text{ or } 43$$

$$n_3 = \left(\frac{8{,}000}{17{,}800}\right) 256 = 115.1 \text{ or } 115$$

$$n_4 = \left(\frac{2{,}800}{17{,}800}\right) 256 = 40.3$$

The number of units allocated to the fourth stratum is greater than the total size of the stratum. Thus every unit in this stratum would be selected ($n_4 = N_4 = 20$) and the sample size for the first three strata recomputed. For these three strata,

$$\sum N_h s_h = 15{,}000$$

$$\sum N_h s_h^2 = 330{,}000$$

Hence,

$$n' = \frac{(15{,}000)^2}{(720^2)(1^2) + 330{,}000} = 265$$

And the allocation of these observations among the three strata would be

$$n_1 = \left(\frac{4{,}000}{15{,}000}\right) 265 = 70.7 \text{ or } 71$$

$$n_2 = \left(\frac{3{,}000}{15{,}000}\right) 265 = 53.0 \text{ or } 53$$

$$n_3 = \left(\frac{8{,}000}{15{,}000}\right) 265 = 141.3 \text{ or } 141$$

Regression Estimators

Regression 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. If we have exact knowledge of the basal area of a stand of timber, the relationship between volume and basal area may help us to improve our estimate of stand volume. The sample data provides information on the volume-basal area relationship which is then applied to the known basal area, giving a volume estimate that may be better or cheaper than would be obtained by sampling volume alone.

Suppose a 100 percent inventory of a 200-acre pine stand indicates a basal area of 84 square feet per acre in trees 3.6 inches in d.b.h. and larger. Assume further that on 20 random plots, each one-fifth acre in size, measurements were made of the basal area (x) and volume (y) per acre.

Basal area per acre (x) (sq. ft.)	*Volume per acre (y) (cu. ft.)*		*Basal area per acre (x) (sq. ft.)*	*Volume per acre (y) (cu. ft.)*
88	1,680		82	1,560
72	1,460		76	1,560
80	1,590		86	1,610
96	1,880		73	1,370
64	1,240		79	1,490
48	1,060		85	1,710
76	1,500		84	1,600
85	1,620		75	1,440
93	1,880			
110	2,140	Total....	1,620	31,860
88	1,840			
80	1,630	Mean......	81	1,593

Some values that will be needed later are

$$n = 20 \qquad \Sigma xy = 2{,}635{,}500$$
$$\Sigma y = 31{,}860 \qquad \Sigma x = 1{,}620$$
$$\bar{y} = 1{,}593 \qquad \bar{x} = 81$$
$$\Sigma y^2 = 51{,}822{,}600 \qquad \Sigma x^2 = 134{,}210$$

$$SS_y = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 51{,}822{,}600 - \frac{(31{,}860)^2}{20} = 1{,}069{,}620$$

$$s_y^2 = \frac{SS_y}{(n-1)} = \frac{1{,}069{,}620}{19} = 56{,}295.79$$

$$SS_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 134{,}210 - \frac{(1{,}620)^2}{20} = 2{,}990$$

$$SP_{xy} = \Sigma xy - \frac{(\Sigma x)\,(\Sigma y)}{n} = 2{,}635{,}500 - \frac{(1{,}620)\,(31{,}860)}{20} = 54{,}840$$

N = total number of fifth-acre plots in the population (= 1,000)

The relationship between y and x may take one of several forms, but here we will assume that it is a straight line. The equation for the line can be estimated from

$$\bar{y}_R = \bar{y} + b\ (X - \bar{x})$$

Where: $\bar{y}_R$ = The mean value of y as estimated from X (a specified value of the variable X).
$\bar{y}$ = The sample mean of y (= 1,593).
$\bar{x}$ = The sample mean of x (= 81).
b = The linear regression coefficient of y on x.

For the linear regression estimator used here, the value of the regression coefficient is estimated by

$$b = \frac{SP_{xy}}{SS_x} = \frac{54{,}840}{2{,}990} = 18.34$$

Thus, the equation would be

$$\begin{aligned}\bar{y}_R &= 1{,}593 + 18.34\ (X - 81)\\ &= 107.46 + 18.34\ X\end{aligned}$$

To estimate the mean volume per acre for the tract we substitute for X the known mean basal area per acre.

$$\bar{y}_R = 107.46 + 18.34\ (84) = 1{,}648 \text{ cubic feet per acre}$$

Standard error.—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 ($s_{y.x}$) which is computed by

$$\begin{aligned}s_{y.x} &= \sqrt{\frac{SS_y - \frac{(SP_{xy})^2}{SS_x}}{(n-2)}}\\ &= \sqrt{\frac{1{,}069{,}620 - \frac{(54{,}840)^2}{2{,}990}}{(20-2)}}\\ &= 59.53\end{aligned}$$

The symbol $s_{y.x}$ bears a strong resemblance to the covariance symbol (s_{yx}) with which it must not be confused.

Having the standard deviation from regression, the standard error of $\bar{y}_R$ is

$$\begin{aligned}s_{\bar{y}_R} &= s_{y.x}\sqrt{\left(\frac{1}{n} + \frac{(X-\bar{x})^2}{SS_x}\right)\left(1 - \frac{n}{N}\right)}\\ &= 59.53\sqrt{\left(\frac{1}{20} + \frac{(84-81)^2}{2{,}990}\right)\left(1 - \frac{20}{1{,}000}\right)}\\ &= 13.57\end{aligned}$$

With such a small sampling fraction $\left(\frac{n}{N} = 0.02\right)$, the finite-population correction $\left(1 - \frac{n}{N}\right)$ could have been ignored, and the standard error would be 13.71.

It is interesting to compare $s_{\bar{y}_R}$ with the standard error that would have been obtained by estimating the mean volume per acre from the y-values only. The estimated mean volume per acre would have been $\bar{y} = 1{,}593$ (compared to 1,648 using the regression estimator). The standard error of this estimate would be

$$s_{\bar{y}} = \sqrt{\frac{s_y^2}{n}\left(1 - \frac{n}{N}\right)}$$

$$= \sqrt{\frac{56{,}295.79}{20}(0.98)}$$

$$= 52.52 \text{ (compared to a standard error of 13.57 with the regression estimator).}$$

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 the next section (Double Sampling), an estimate of the population mean can often be substituted.

Another condition is that the relationship of y to x must be reasonably close to a straight line within the range of x values for which y will be estimated. If the relationship departs very greatly from a straight line, our estimate of the mean value of y will not be reliable. Often a curvilinear function is more appropriate.

A third condition is that the variance of y about its mean should be the same at all levels of x. This condition is difficult to evaluate with the amount of data usually available. Ordinarily the question is answered from our knowledge of the population or by making special studies of the variability of y. If we know the way in which the variance changes with changes in the level of x a weighted regression procedure may be used.

Thus, 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 *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. The ratio estimate ($\bar{y}_R$) of mean y is

$$\bar{y}_R = \hat{R}\,X$$

Where: R = The ratio of means obtained from the sample

$$= \frac{\bar{y}}{\bar{x}} \text{ or } \frac{\Sigma y}{\Sigma x}$$

X = The known population mean of x.

The standard error of this estimate can be reasonably approximated for large samples by

$$s_{\bar{y}_R} = \sqrt{\left(\frac{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}}{n}\right)\left(1 - \frac{n}{N}\right)}$$

Where: s_y^2 = The estimated variance of y.
s_x^2 = The estimated variance of x.
s_{xy} = The estimated covariance of x and y.

It is difficult to say when a sample is large enough for the standard error formula to be reliable, but Cochran (see References, p. 78) has suggested that n must be greater than 30 and also large enough so that the ratios $s_{\bar{y}}/\bar{y}$ and $s_{\bar{x}}/\bar{x}$ are both less than 0.1.

To illustrate the computations, assume that for a population of $N = 400$ units, the population mean of x is known to be 62 and that from this population a sample of $n = 10$ units is selected. The y and x values for these 10 units are found to be

Observation	y_i	x_i	*Observation*	y_i	x_i
1	8	62	8	11	96
2	13	81	9	5	36
3	5	40	10	12	70
4	6	46			
5	19	123	Total	96	680
6	9	74			
7	8	52	Mean	9.6	68

From this sample the ratio-of-means is

$$\hat{R} = \frac{9.6}{68} = 0.141$$

The ratio-of-means estimator is then

$$\bar{y}_R = \hat{R}X = 0.141\,(62) = 8.742$$

To compute the standard error of the mean we will need the variances of y and x and also the covariance. These values are computed by the standard formulae for a simple random sample. Thus,

$$s_y^2 = \frac{(8^2 + 13^2 + \ldots + 12^2) - \frac{(96)^2}{10}}{(10 - 1)} = 18.7111$$

$$s_x^2 = \frac{(62^2 + 81^2 + \ldots + 70^2) - \frac{(680)^2}{10}}{(10 - 1)} = 733.5556$$

$$s_{xy} = \frac{(8)\,(62) + (13)\,(81) + \ldots + (12)\,(70) - \frac{(96)\,(680)}{10}}{(10 - 1)}$$

$$= 110.2222$$

Substituting these values in the formula for the standard error of the mean gives

$$s_{\bar{y}_R} = \sqrt{\left(\frac{(18.7111) + (0.141^2)(733.5556) - 2(0.141)(110.2222)}{10}\right)\left(1 - \frac{10}{400}\right)}$$

$$= \sqrt{.215690}$$

$$= 0.464$$

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

The *mean-of-ratios estimator* is appropriate when the relation of y to x is in the form of a straight line passing through the origin *and* the standard deviation of y at a given level of x is proportional to x (rather than to $\sqrt{x}$). The ratio (r_i) of y_i to x_i is computed for each pair of sample observations. Then the estimated mean of y for the population is

$$\bar{y}_R = \hat{R}X$$

Where: $\hat{R}$ = the mean of the individual ratios (r_i), i.e.,

$$\hat{R} = \frac{\sum_{i=1}^{n} r_i}{n}$$

To compute the standard error of this estimate we must first obtain a measure (s_r^2) of the variability of the individual ratios (r_i) about their mean.

$$s_r^2 = \frac{\sum_{i=1}^{n} r_i^2 - \frac{\left(\sum_{i=1}^{n} r_i\right)^2}{n}}{(n-1)}$$

The standard error for the mean-of-ratios estimator of mean y is then

$$s_{\bar{y}_R} = X\sqrt{\frac{s_r^2}{n}\left(1 - \frac{n}{N}\right)}$$

Suppose that a set of $n = 10$ observations is taken from a population of $N = 100$ units having a mean x value of 40:

Observation	y_i	x_i	r_i
1	36	18	2.00
2	95	48	1.98
3	108	46	2.35
4	172	74	2.32
5	126	58	2.17
6	58	26	2.23
7	123	60	2.05
8	98	51	1.92
9	54	25	2.16
10	14	7	2.00
Total			21.18

The sample mean-of-ratios is

$$\hat{R} = \frac{21.18}{10} = 2.118$$

And this is used to obtain the mean-of-ratios estimator

$$\bar{y}_R = \hat{R}X = 2.118(40) = 84.72$$

The variance of the individual ratios is

$$s_r^2 = \frac{(2.00^2 + 1.98^2 + \ldots + 2.00^2) - \frac{(21.18)^2}{10}}{(10 - 1)} = 0.022484$$

Thus, the standard error of the mean-of-ratios estimator is

$$s_{\bar{y}_R} = 40\sqrt{\frac{0.022484}{10}\left(1 - \frac{10}{100}\right)}$$
$$= 1.799$$

Numerous other forms of ratio estimators are possible, but the above three are the most common. Less common forms involve fitting some curvilinear function for the relationship of y to x, or fitting multiple regressions when information is available on more than one supplementary variable.

Warning! The forester who is not sure of his knowledge of regression 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 tricky. The two 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). 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 estimate 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.

Double Sampling

Double sampling was devised to permit the use of regression estimators when the population mean or total of the supplementary variable is unknown. A large sample is taken in order to obtain a good estimate of the mean or total for the supplementary variable (x). On a subsample of the units in this large sample, the y values are also measured to provide an estimate of the relationship of y to x. The large sample mean or total of x is then applied to the fitted relationship to obtain an estimate of the population mean or total of y.

Updating a forest inventory is one application of double sampling. Suppose that in 1950 a sample of 200 quarter-acre plots in an 800-acre forest showed a mean volume of 372 cubic feet per plot (1,488 cubic feet per acre). A subsample of 40 plots, selected at random from the 200 plots, was marked for remeasurement in 1955. The relationship of the 1955 volume to the 1950 volume as determined from the subsample was applied to the 1950 volume to obtain a regression estimate of the 1955 volume.

The subsample was as follows:

1955 volume (y)	*1950 volume* (x)		*1955 volume* (y)	*1950 volume* (x)
370	280		550	430
290	240		550	460
520	410		520	400
490	360		420	390
530	390		490	340
330	220		500	420
310	270		610	470
400	340		460	350
450	360		430	340
430	360		510	380
460	400		450	370
480	380		380	300
430	350		430	290
500	390		460	340
640	480		490	370
660	520		560	440
490	400		580	480
510	430		540	420
270	230			
380	270	Total ..	18,820	14,790
420	330			
530	390	Mean ..	470.50	369.75

$\Sigma\, y^2 = 9{,}157{,}400$

$\Sigma\, x^2 = 5{,}661{,}300$

$\Sigma\, xy = 7{,}186{,}300$

A plotting of the 40 pairs of plot values on coordinate paper suggested that the variability of y was the same at all levels of x and that the relationship of y to x was linear. The estimator selected on the basis of this information was the linear regression $\bar{y}_{Rd} = a + bX$. Values needed to compute the linear-regression estimate and its standard error were as follows:

Large-sample data (indicated by the subscript 1):

n_1 = Number of observations in large sample = 200

N = Number of sample units in population = 3,200

$\bar{x}_1$ = Large sample mean of $x \doteq 372$

Small-sample data (indicated by the subscript 2):

n_2 = Number of observations in subsample = 40

$\bar{y}_2$ = Small sample mean of y = 470.50

$\bar{x}_2$ = Small sample mean of x = 369.75

$$SS_y = \left(\Sigma y^2 - \frac{(\Sigma y)^2}{n_2}\right) = \left(9{,}157{,}400 - \frac{(18{,}820)^2}{40}\right) = 302{,}590.0$$

$$SS_x = \left(\Sigma x^2 - \frac{(\Sigma x)^2}{n_2}\right) = \left(5{,}661{,}300 - \frac{(14{,}790)^2}{40}\right) = 192{,}697.5$$

$$SP_{xy} = \left(\Sigma xy - \frac{(\Sigma x)\,(\Sigma y)}{n_2}\right) = \left(7{,}186{,}300 - \frac{(18{,}820)\,(14{,}790)}{40}\right)$$

$$= \ 227{,}605.0$$

$$s_y^2 = \frac{SS_y}{(n_2 - 1)} = \frac{302{,}590}{40 - 1} = 7{,}758.72$$

The regression coefficient (b) and the squared standard deviation from regression ($s_{y.x}$) are

$$b = \frac{SP_{xy}}{SS_x} = \frac{227{,}605.0}{192{,}697.5} = 1.18$$

$$s_{y.x}^2 = \frac{\left(SS_y - \dfrac{(SP_{xy})^2}{SS_x}\right)}{(n_2 - 2)} = \frac{\left(302{,}590.0 - \dfrac{(227{,}605.0)^2}{192{,}697.5}\right)}{40 - 2} = 888.2617$$

And the regression equation is

$$\begin{aligned} \bar{y}_{Rd} &= \bar{y}_2 + b\ (X - \bar{x}_2) \\ &= 470.50 + 1.18\ (X - 369.75) \\ &= 34.2 + 1.18X \end{aligned}$$

Substituting the 1950 mean volume (372 cubic feet) for X gives the regression estimate of the 1955 volume.

$$\bar{y}_{Rd} = 34.2 + 1.18(372) = 473.16 \text{ cubic feet per plot}$$

Standard error.—The standard error of $\bar{y}_{Rd}$ when the linear-regression estimator is used in double sampling is

$$\begin{aligned} s_{\bar{y}_{Rd}} &= \sqrt{s_{y.x}^2\left(\frac{1}{n_2} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{SS_x}\right)\left(1 - \frac{n_2}{n_1}\right) + \frac{s_y^2}{n_1}\left(1 - \frac{n_1}{N}\right)} \\ &= \sqrt{888.2617\left(\frac{1}{40} + \frac{(372.00 - 369.75)^2}{192{,}697.5}\right)\left(1 - \frac{40}{200}\right)} \\ &\quad \overline{+ \frac{7{,}758.72}{200}\left(1 - \frac{200}{3{,}200}\right)} \\ &= 7.36 \text{ cubic feet} \end{aligned}$$

Had the 1955 volume been estimated from the 40 plots without taking advantage of the relationship of y to x, the estimated mean would have been

$$\bar{y} = \frac{18{,}820}{40} = 470.50 \text{ cubic feet (instead of 473.16)}$$

The standard error of $\bar{y}$ would have been

$$s_{\bar{y}} = \sqrt{\frac{s_y^2}{n_2}\left(1 - \frac{n_2}{N}\right)}$$

$$= \sqrt{\frac{7{,}758.72}{40}\left(1 - \frac{40}{3{,}200}\right)}$$

$$= 13.84 \text{ cubic feet (compared to 7.36)}$$

Double sampling with other regression estimators.—If the mean-of-ratios estimate is deemed appropriate, the individual ratios ($r_i = y_i/x_i$) are computed for the n_2 observations of the subsample. The mean of ratios estimate is then

$$\bar{y}_{Rd} = \hat{R}\bar{x}_1$$

with standard error

$$s_{\bar{y}_{Rd}} = \sqrt{\bar{x}_1^2\left(\frac{s_r^2}{n_2}\right)\left(1 - \frac{n_2}{n_1}\right) + \frac{s_y^2}{n_1}\left(1 - \frac{n_1}{N}\right)}$$

Where: $\hat{R} = \dfrac{\Sigma r_i}{n_2}$

$\bar{x}_1$ = Mean x for the large sample of n_1 observations

s_r^2 = Variance of r for the subsample

$$= \frac{\Sigma r_i^2 - \dfrac{(\Sigma r_i)^2}{n_2}}{n_2 - 1}$$

The ratio-of-means estimate, when appropriate, is

$$\bar{y}_{Rd} = \hat{R}\bar{x}_1$$

with standard error

$$s_{\bar{y}_{Rd}} = \sqrt{\left(1 - \frac{n_2}{n_1}\right)\left(\frac{\bar{x}_1}{\bar{x}_2}\right)^2\left(\frac{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}}{n_2}\right) + \frac{s_y^2}{n_1}\left(1 - \frac{n_1}{N}\right)}$$

Where: $\hat{R} = \bar{y}_2/\bar{x}_2$

s_y^2 = Variance of y in the subsample.

s_x^2 = Variance of x in the subsample.

s_{yx} = Covariance of y and x in the subsample.

Sampling When Units are Unequal in Size (Including PPS Sampling)

Sampling units of unequal size are common in forestry. Plantations, farms, woodlots, counties, and sawmills are just a few of the natural units that vary in size. Designing and analyzing surveys involving unequal-sized units can be quite tricky. Two examples will be used to illustrate the problem and some of the possible solutions. They also illustrate the very important fact that no single method is best for all cases and that designing an efficient survey requires considerable skill and caution.

Example No. 1.—As a first example, suppose that we want to estimate the mean milling cost per thousand board feet of lumber at southern pine sawmills in a given area. Available for planning the survey is a list of the 816 sawmills in the area and the daily capacity of each. The cost information is to be obtained by personal interview.

In sampling, as in most other endeavors, the simplest approach that will do the job is the best; complex procedures should be used only when they offer definite advantages. On this principle we might first consider taking a simple random sample of the mills, obtaining the cost per thousand at each, and computing the arithmetic average of these values. Most foresters would reject this procedure, and rightly so. The design would give the same weight to the cost for a mill producing 8,000 feet per day as to the cost for one cutting 50,000 feet per day. As a result, one thousand feet at the small mill would have a larger representation in the final average than the same volume at the large mill, and because cost per thousand is undoubtedly related to mill capacity, the estimate would be biased.

An alternative that would give more weight to the large mills would be to take a random sample of the mills, obtain the total milling cost (y_i) and the total production in MBF (x_i) at each, and then use the ratio-of-means estimator:

$$\text{Mean cost per MBF} = \frac{\text{Total cost at all sampled mills}}{\text{Total production at all sampled mills}} = \frac{\Sigma y_i}{\Sigma x_i}$$

This must also be rejected on the grounds of bias. The ratio-of-means estimator is unbiased only if the ratio of y to x is the same at all levels of x. In this example, a constant ratio of y to x means that the milling cost per thousand is the same regardless of mill size—an unlikely situation.

An unbiased procedure and one that would be appropriate in this situation is sampling with probability proportional to size (known as pps sampling). The value to be observed on each sample unit would be the milling cost per thousand board feet of lumber. Selection of the units with probability proportional to size is easily accomplished.

First, a list is made of all of the mills along with their daily capacities and the cumulative sum of capacities.

Mill No.	*Daily capacity (MBF)*	*Cumulative sum*
1	10	10
2	27	37
3	8	45
4	12	57
–	–	–
–	–	–
–	–	–
–	–	–
814	13	12,210
815	21	12,231
816	11	12,242
	12,242	

Next, numbers varying in size from 1 up to the cumulative sum for the last mill on the list (12,242) are selected from a table of random digits. A particular mill is included in the sample when a number is drawn which is equal to or less than the cumulative sum for that mill and greater than the cumulative sum for the preceding mill. Thus, given a random number of 49 we would select mill number 4; for 37 we would select mill number 2; for 12,238 we would select mill number 816. An important point is that sampling must be with replacement (i.e., a given mill may appear in the sample more than once); otherwise, sampling will not be proportional to size.

After the sample units have been selected and the unit values (y_i = milling cost per thousand) obtained, the mean cost per thousand and the standard error of the mean are computed in the same manner as for simple random sampling with replacement.

Given the following ten observations:

Mill	*Milling cost per MBF (dollars)*	*Mill*	*Milling cost per MBF (dollars)*
73	12	329	11
541	13	804	17
126	18	126	18
134	14	427	12
423	16		
703	21	Total	152

The estimated mean is

$$\bar{y} = \frac{152}{10} = 15.2 \text{ dollars per thousand}$$

The standard error of the mean is

$$s_{\bar{y}} = \sqrt{\frac{s_y^2}{n}} = \sqrt{\frac{\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}}{n(n-1)}}$$

$$= \sqrt{\frac{2{,}408 - \frac{(152)^2}{10}}{10(9)}}$$

$$= 1.04$$

Another alternative is to group mills of similar size into strata and use stratified random sampling. If the cost per thousand is related to mill size, this procedure may be slightly biased unless all mills in a given stratum are of the same size. With only a small within-stratum spread in mill size, the bias will usually be trivial.

A further refinement would be to group mills of similar size and use stratified random sampling with pps sampling of units within strata.

Example No. 2.—Now, consider the problem of estimating the total daily production of chippable waste at these mills. Assume again that we have a list of the mills and their daily capacities.

We might first consider a simple random sample of the mills with the unit observation being the mean daily production of chippable waste at the selected mills. The arithmetic average of these observations multiplied by the total number of mills would give an estimate of the total daily production of chippable waste by all mills. This estimate would be completely unbiased. However, because the mills vary greatly in daily capacity and because total waste production is closely related to total lumber production, there will be a large variation in chippable waste from unit to unit. This means that the variance among units will be large and that many observations may be needed to obtain an estimate of the desired precision. The simple random sample, though unbiased, would probably be rejected because of its low precision.

The ratio-of-means estimator is a second alternative. In this design a simple random sample would be selected and for each mill included in the sample we would observe the mean daily production of chippable waste (y_i) and the mean daily capacity of the mill in MBF (x_i). The ratio of means

$$\hat{R} = \frac{\Sigma y_i}{\Sigma x_i}$$

would give an estimate of the mean waste production per MBF, and this ratio multiplied by the total capacity of all mills would estimate the total daily production of chippable waste. It has been pointed out that the ratio-of-means estimator is unbiased if the ratio of y to x is the same at all levels of x. Studies have shown that although the ratio of waste to lumber production varies with log size, it is not closely related to mill size—hence the bias, if any, in the ratio-of-means estimator would be small. Past experience suggests that the variance of the estimate will also be small, making it preferable to the simple arithmetic average previously discussed. Note that this is a case where a slightly biased estimator of high precision might be more suitable than an unbiased estimator of low precision.

Here again, pps sampling would merit consideration. It would give unbiased estimates of moderately good precision. Stratified sampling with units grouped according to size is another possibility as is the combination of stratification with pps sampling within strata. Among the acceptable alternatives no blanket recommendation is possible. The best choice depends on many factors, chief among them being the form and closeness of the relationship between chippable waste (y_i) and mill capacity (x_i).

Two-Stage Sampling

In some forest sampling, locating and getting to a sampling unit is expensive, while measurement of the unit is relatively cheap. It seems logical in these circumstances to make measurements on two or three units at or near each location. This is called two-stage sampling, the first stage being the selection of locations, and the second stage being the selection of units at these locations. The advantage of two-stage sampling is that it may yield estimates of a given precision at a cost lower than that of a completely random sample.

To illustrate the situation and the methods, consider a landowner whose 60,000 acres of timberland are subdivided into square blocks of 40 acres with permanent markers at the four corners of each block. A sample survey is to be made of the tract in order to estimate the mean sawtimber volume per acre. Sample units will be square quarter-acre plots. These plots will be located on the ground by measurements made with reference to one of the corners of the 40-acre blocks.

Travel and surveying time to a block corner are quite high, hence it seems logical, once the block corner is located, to find and measure several plots in that block. Thus, the sampling scheme would consist of making a random selection of n blocks and then randomly selecting m plots within each of the selected blocks. In sampling language, the 40-acre blocks would be called primary sampling units (primaries) and the quarter-acre plots secondary sampling units (secondaries).

If y_{ij} designates the volume of the j^{th} sampled plot ($j = 1 \ldots m$) on the i^{th} sampled block, the estimated mean volume *per plot* (symbolized in two-stage sampling by $\bar{\bar{y}}$) is

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}}{mn}$$

The standard error of the estimated mean is

$$s_{\bar{\bar{y}}} = \sqrt{\frac{1}{mn}\left[s_B^2\left(1 - \frac{n}{N}\right) + \frac{n s_W^2}{N}\left(1 - \frac{m}{M}\right)\right]}$$

Where: n = Number of primaries sampled.

N = Total number of primaries in the population.

m = Number of secondaries sampled in each of the primaries selected for sampling.

M = Total number of secondaries in each primary.

s_B^2 = Sample variance between primaries when sampled by m secondaries per primary (computation procedure given below).

s_W^2 = Sample variance among secondaries within primaries (computation procedure given below).

The terms s_B^2 and s_W^2 are computed from the equations

$$s_B^2 = \frac{\dfrac{\sum\limits_{i=1}^{n}\left(\sum\limits_{j=1}^{m} y_{ij}\right)^2}{m} - \dfrac{\left(\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} y_{ij}\right)^2}{mn}}{(n-1)}$$

$$s_W^2 = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} y_{ij}^2 - \dfrac{\sum\limits_{i=1}^{n}\left(\sum\limits_{j=1}^{m} y_{ij}\right)^2}{m}}{n(m-1)}$$

Since y_{ij} is the observed value of a secondary unit, $\sum\limits_{j=1}^{m} y_{ij}$ is the total of all secondary units observed in the i^{th} primary (or the primary total), and $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} y_{ij}$ is the grand total of all sampled secondaries. Hence, the above equations, expressed in words, are

$$s_B^2 = \frac{\left(\dfrac{\sum\limits^{n} (\text{Primary totals}^2)}{\text{No. of secondaries sampled per primary}}\right) - \left(\dfrac{\left[\sum\limits^{mn} (\text{Secondaries})\right]^2}{\text{Total no. of secondaries sampled}}\right)}{(n-1)}$$

$$s_W^2 = \frac{\sum\limits^{mn} (\text{Secondaries}^2) - \left(\dfrac{\sum\limits^{n} (\text{Primary totals}^2)}{\text{No. of secondaries sampled per primary}}\right)}{n(m-1)}$$

Readers familiar with analysis of variance procedures will recognize s_B^2 and s_W^2 as the mean square between and within primaries respectively.

The computations are not so difficult as the notation might suggest. Suppose we had sampled $m = 3$ quarter-acre plots (secondaries) within each of $n = 4$ blocks (primaries) and obtained the following data:

Block (primary)	*Plot (secondary)*	*Secondary values* *(cubic feet)*	*Primary totals* *(cubic feet)*
1	1	147	
	2	180	
	3	206	533
2	1	312	
	2	265	
	3	300	877
3	1	220	
	2	280	
	3	210	710
4	1	250	
	2	232	
	3	185	667
		2,787	2,787

The estimated mean per plot is

$$\bar{y} = \frac{(147 + 180 + \ldots + 185)}{(3)\,(4)} = \frac{2{,}787}{12} = 232.25 \text{ cubic feet per plot.}$$

To get the standard error of $\bar{y}$ we first compute $s_B{}^2$ and $s_W{}^2$.

$$s_B{}^2 = \frac{\dfrac{(533^2 + \ldots + 667^2)}{3} - \dfrac{(2{,}787)^2}{(3)\,(4)}}{(4-1)}$$

$$= \frac{667{,}402.3333 - 647{,}280.7500}{3}$$

$$= 6{,}707.1944$$

$$s_W{}^2 = \frac{\left(147^2 + 180^2 + \ldots + 185^2\right) - \dfrac{(533^2 + \ldots + 667^2)}{3}}{(4)\,(3-1)}$$

$$= \frac{675{,}463.0000 - 667{,}402.3333}{8}$$

$$= 1{,}007.5833$$

Since the total number of 40-acre blocks in the 60,000 acres is $N = 1{,}500$ and the total number of quarter-acre plots in each 40-acre block is $M = 160$, the estimated standard error of the mean is

$$s_{\bar{y}} = \sqrt{\frac{1}{(3)\,(4)}\left[6{,}707.1944\left(1 - \frac{4}{1{,}500}\right) + \frac{(4)\,(1{,}007.5833)}{1{,}500}\left(1 - \frac{3}{160}\right)\right]}$$

$$= \sqrt{\frac{1}{12}\,[6{,}689.3085 + 2.6365]}$$

$$= 23.61$$

The estimated mean *per plot* is 232.25 cubic feet. The standard error of this estimate is 23.61 cubic feet. As the plots are one-quarter acre in size, the estimated mean volume per acre is $4(\bar{y}) = 929$ cubic feet. The standard error of the mean volume per acre is $4(s_{\bar{y}}) = 94.44$.

An estimate of the total volume and its standard error can be obtained either from the mean per plot or mean per acre volumes and their standard errors. The mean per plot is 232.25 ± 23.61. To expand this to the total, each figure must be multiplied by the number of quarter-acre plots in the entire tract (= 240,000) ; the estimated total is

$$55{,}740{,}000 \pm 5{,}666{,}400.$$

The mean per acre is 929 ± 94.44. To expand this, each figure must be multiplied by the total number of acres in the tract (= 60,000). Thus, the estimated total is

$$55{,}740{,}000 \pm 5{,}666{,}400 \text{ as before.}$$

Small sampling fractions.—If the number of primary units sampled (n) is a small fraction of the total number of primary units (N), the standard error formula simplifies to

$$s_{\bar{y}} = \sqrt{\frac{s_B^2}{mn}}$$

This reduced formula is usually applied where the ratio n/N is less than 0.01. In the example above, the sampling fraction for primaries was 4/1,500, so we could very well have used the short formula. The estimated standard error would have been

$$s_{\bar{y}} = \sqrt{\frac{6{,}707.1944}{3(4)}} = \sqrt{558.9329}$$
$$= 23.64 \text{ (instead of 23.61 by the longer formula).}$$

When n/N is fairly large but the number of secondaries (m) sampled in each selected primary is only a small fraction of the total number of secondaries (M) in each primary, the standard error formula would be

$$s_{\bar{y}} = \sqrt{\frac{1}{mn}\left[s_B^2\left(1 - \frac{n}{N}\right) + \frac{ns_W^2}{N}\right]}$$

Sample size for two-stage sampling.—For a fixed number of sample observations, two-stage sampling is usually *less* precise than simple random sampling. The advantage of the method is that by reducing the cost per observation it permits us to obtain the desired precision at a lower cost.

Usually the precision and cost both increase as the number of primaries is increased and the number of secondaries (m) per sampled primary is decreased. The cost may be reduced by taking fewer primaries and more secondaries per primary, but precision usually suffers. This suggests that there is a number (m) of secondaries per primary that will be optimum from the standpoint of giving the greatest precision for a given amount of money. The value of m that is optimum depends on the nature of the population variability between primaries and among secondaries within primaries, and on the relationship between the cost per primary and the added cost per secondary.

The population variability between primaries is symbolized by σ_I^2 and the variability within primaries by σ_{II}^2. Note that these are population values, not sample values. Occasionally we will have some knowledge of σ_I^2 and σ_{II}^2 from previous work with the population. More often, it will be necessary to take a preliminary sample to estimate the population variabilities. From this presample, we compute s_B^2 and s_W^2 according to the formulae in the discussion of the error of a two-stage sample. Then our estimates of the population variability within and between primaries are

$$\sigma_{II}^2 = s_W^2$$

$$\sigma_I^2 = \frac{s_B^2 - s_W^2}{m}$$

The cost of locating and establishing a primary unit (not counting overhead costs) is symbolized by c_p. The additional cost of getting to and measuring a secondary unit after the primary has been located is symbolized by c_s.

Given the necessary cost and variance information, we can estimate the optimum size of m (say m_o) by

$$m_o = \sqrt{\left(\frac{\sigma_{II}^2}{\sigma_I^2}\right)\left(\frac{c_p}{c_s}\right)}$$

If m_o is greater than the number of secondaries per primary (M), the formula value is ignored and m_o is set equal to M.

Once m_o has been estimated, the number of primary units (with m_o secondaries per primary) needed to estimate the mean with a specified standard error (D) is

$$n = \frac{\left(\sigma_I^2 + \frac{\sigma_{II}^2}{m_o}\right)}{D^2 + \frac{1}{N}\left(\sigma_I^2 + \frac{\sigma_{II}^2}{M}\right)}$$

Where: N = Total number of primaries in the population.
M = Total number of secondaries per primary.

Numerical example.—Suppose that we wish to estimate a population mean with a standard error of 10 percent or less. We have defined the population as being composed of $N = 1{,}000$ primaries with $M = 100$ secondaries per primary.

As we know nothing of the variability between or within these primaries nor about the costs, we take a preliminary sample consisting of eight primaries with two secondaries per primary. Results are as follows:

Data from preliminary survey

Primary	*Observed values of secondaries*		*Primary total*
1	34	42	76
2	36	17	53
3	41	56	97
4	62	40	102
5	82	94	176
6	16	38	54
7	22	41	63
8	93	50	143
		Total =	764

From this preliminary sample, we compute

$$s_B^2 = 981.8571$$

$$s_W^2 = 248.2500$$

$$\bar{y} = \frac{764}{16} = 47.75$$

Therefore, the estimates of the population variances between and within primaries are

$$\sigma_{II}^2 = s_W^2 = 248.25$$

$$\sigma_I^2 = \frac{s_B^2 - s_W^2}{m} = \frac{981.8571 - 248.2500}{2} = 366.8036$$

Assume also that the preliminary sample yields the following cost estimates:

$$c_p = \$14.00$$

$$c_s = \$\ 1.20$$

Then our estimate of the optimum number of secondaries to be observed in each primary is

$$\begin{aligned} m_o &= \sqrt{\left(\frac{\sigma_{II}^2}{\sigma_I^2}\right)\left(\frac{c_p}{c_s}\right)} \\ &= \sqrt{\left(\frac{248.25}{366.8036}\right)\left(\frac{14.00}{1.20}\right)} \\ &= \sqrt{(.6768)\,(11.6667)} \\ &= \sqrt{7.8960} \\ &= 2.8 \end{aligned}$$

Since we can't observe a fraction of a unit, we must now decide whether to take two or three secondaries per primary. To do this, we estimate the number of primaries needed for an m of 2 and for an m of 3, compute the cost of the two alternatives and select the less expensive one.

Our preliminary sample gave an estimate of the mean of 47.75 and, since we have specified a standard error of 10 percent, this means we want $D = (0.10)(47.75) = 4.775$ or 4.8.

If $m = 2$, the number of primaries needed for the desired precision would be

$$\begin{aligned} n &= \frac{\left(\sigma_I^2 + \frac{\sigma_{II}^2}{m_o}\right)}{D^2 + \frac{1}{N}\left(\sigma_I^2 + \frac{\sigma_{II}^2}{M}\right)} \\ &= \frac{\left(366.8036 + \frac{248.25}{2}\right)}{(4.8)^2 + \frac{1}{1{,}000}\left(366.8036 + \frac{248.25}{100}\right)} \\ &= \frac{490.9286}{23.4093} \\ &= 20.97 \\ \text{or, } n &= 21 \end{aligned}$$

There will be 21 primaries at a cost of \$14 each and $2(21) = 42$ secondaries at a cost of \$1.20 each, so that the total survey cost (exclusive of overhead) will be \$344.40.

If $m = 3$, the number of primaries will be

$$\begin{aligned} n &= \frac{\left(366.8036 + \frac{248.25}{3}\right)}{(4.8)^2 + \frac{1}{1{,}000}\left(366.8036 + \frac{248.25}{100}\right)} \\ &= \frac{449.5536}{23.4093} = 19.20 \\ \text{or, } n &= 20 \end{aligned}$$

The cost of this survey will be

$$20(14.00) + 60(1.20) = 352.00$$

As the first alternative gives the desired precision at a lower cost, we would sample $n = 21$ primaries and $m = 2$ secondaries per primary.

Systematic arrangement of secondaries.—Though the potential economy of two-stage sampling has been apparent and appealing to foresters, they have displayed a reluctance to select secondary units at random. Primary sampling points may be selected at random, but at each point the secondaries will often be arranged in a set pattern. This is not two-stage sampling in the sense that we have been using the term, though it may result in similar increases in sampling efficiency. It might be called cluster sampling, the cluster being the group of secondaries at each location. The unit of observation then is not the individual plot but the entire cluster. The unit value is the mean or total for the cluster. Estimates and their errors are computed by the formulae that apply to the method of selecting the cluster locations.

Within each primary the clusters should be selected so that every secondary has a chance of appearing in the sample. If certain portions of the primaries are systematically excluded, bias may result.

Two-Stage Sampling With Unequal-Sized Primaries

The two-stage method of the previous chapter gives the same weight to all primaries. This hardly seems logical if the primaries vary greatly in size. It would, for example, give the same weight to a 10,000-acre tract as to a 40-acre tract. There are several modified methods of two-stage sampling which take primary size into account.

Stratified two-stage sampling.—One approach is to group equal-sized primaries into strata and apply the standard two-stage methods and computations within each stratum. Population estimates are made by combining the individual stratum estimates according to the stratified sampling formulae. This is a very good design if the size of each primary is known and the number of strata is not too large. If the number of primaries is small, it may even be feasible to regard each primary as a stratum and use regular single-stage stratified sampling.

Selecting primaries with probability proportional to size.—Another possibility is to select primaries with probability proportional to size (pps) and secondaries within primaries with equal probability. Selection of primaries must be with replacement, but secondaries can be selected without replacement. A new set of secondaries should be drawn each time that a given primary is selected so that a secondary that was selected during one sampling may again be selected during some subsequent sampling of that primary.

After the observations have been made, the sample mean ($\bar{y}_i$) is computed for each of the n primaries included in the sample. These primary means are then used to compute an estimate of the population mean by

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

The standard error of the mean is

$$s_{\bar{y}} = \sqrt{\frac{\sum_{i=1}^{n} \bar{y}_i^2 - \frac{\left(\sum_{i=1}^{n} \bar{y}_i\right)^2}{n}}{n(n-1)}}$$

If only one secondary is selected in each selected primary, this procedure becomes identical to simple random sampling.

If there is any relationship between the primary size and its mean, pps sampling may give estimates of low precision. The precision can be improved by combining stratified two-stage sampling and pps selection of primaries. Primaries of similar size are grouped into strata and within each stratum selection of primaries is made with probability proportional to size. Strata means and variance are computed by the formulae for two-stage sampling with pps selection of primaries.

Selection of primaries with equal probability.—The procedures that have been discussed so far require reasonably accurate information about the size of each primary in the population—information that is often lacking. An alternative technique requires knowledge only of the size of the primaries actually included in the sample and of the total number of primaries in the population. The method involves selection of n primaries and m_i secondaries within the i^{th} selected primary. At each level, sampling is with equal probability and without replacement. The number of secondaries sampled (m_i) may vary or remain constant. The sample primary mean $(\bar{y}_i)$ is computed for each selected primary and from these the population mean is estimated as

$$\bar{y} = \frac{\sum_{i=1}^{n} (M_i\bar{y}_i)}{\left(\sum_{i=1}^{n} M_i\right)}$$

where: n = Number of primaries sampled.
$\bar{y}_i$ = Mean per secondary in the i^{th} sampled primary.
M_i = Total number of secondary units in the i^{th} sampled primary (this can be an actual or a relative measure of size).

The standard error of this estimate is

$$s_{\bar{y}} = \bar{y}\sqrt{\frac{n}{n-1}\left(\frac{\Sigma M_i^2}{(\Sigma M_i)^2} + \frac{\Sigma T_i^2}{(\Sigma T_i)^2} - \frac{2\,\Sigma M_iT_i}{(\Sigma M_i)(\Sigma T_i)}\right)\left(1 - \frac{n}{N}\right)}$$

where: n = Number of primaries sampled.
N = Total number of primaries.
$T_i = (M_i\bar{y}_i)$

For an illustration of the computations, suppose that we wished to estimate the mean board-foot volume on a population of 426 woodlots. Four woodlots (primaries) are selected at random, and within each woodlot the board-foot volume is measured on two randomly selected one-fifth-acre plots. For each woodlot selected, the acreage is also determined. Since one-fifth-acre plots were used, the value of M_i for the i^{th} woodlot will be 5 times its acreage. Assume the observed values are as follows:

Sampled woodlot	*Plot values*		*Primary means* ($\bar{y}_i$)	*Woodlot acreage*	M_i	$M_i\bar{y}_i = T_i$
	Bd. ft.	*Bd. ft.*				
1	620	740	680	110	550	374,000
2	585	475	530	26	130	68,900
3	590	730	660	54	270	178,200
4	960	820	890	60	300	267,000
					1,250	888,100

Then

$$\bar{y} = \frac{\Sigma(M_i\bar{y}_i)}{(\Sigma M_i)} = \frac{888{,}100}{1{,}250} = 710.48 \text{ board feet per fifth-acre plot.}$$

The values needed to compute the standard error are

$$\Sigma M_i^2 = 482{,}300 \qquad \Sigma T_i^2 = 247{,}667{,}450{,}000$$

$$\Sigma M_i T_i = 342{,}871{,}000$$

$$(\Sigma M_i)^2 = 1{,}562{,}500 \qquad (\Sigma T_i)^2 = 788{,}721{,}610{,}000$$

$$(\Sigma M_i)(\Sigma T_i) = 1{,}110{,}125{,}000$$

Hence,

$$\begin{aligned} s_{\bar{y}} &= 710.48 \sqrt{\left(\frac{4}{3}\right)\left(\frac{482{,}300}{1{,}562{,}500} + \frac{247{,}667{,}450{,}000}{788{,}721{,}610{,}000} - \frac{(2)\,(342{,}871{,}000)}{1{,}110{,}125{,000}}\right.} \quad \left(\begin{matrix}\text{fpc}\\ \text{ignored}\end{matrix}\right) \\ &= 710.48\sqrt{0.00662295} \\ &= 57.82 \text{ board feet.} \end{aligned}$$

This estimate of the mean will be slightly biased if there is any relationship between the primary size and the mean per unit in that primary. The bias is generally not serious for large samples (more than 30 primaries).

An unbiased equal-probability estimator.—If the bias incurred by use of the above estimator is expected to be large, an unbiased estimate can be obtained. In addition to the information required for the biased procedure, we must also know the total number of secondaries (M) in the population. As in the case of the biased estimator, n primaries are selected with equal probability and within each primary m_i secondaries are observed. The mean per unit ($\bar{y}_i$) is computed for each primary and used to estimate the population mean

$$\bar{\bar{y}} = \frac{N}{nM}\sum_{i=1}^{n}(M_i\bar{y}_i)$$

The standard error of the mean is

$$s_{\bar{\bar{y}}} = \frac{N}{M}\sqrt{\frac{\Sigma(M_i\bar{y}_i)^2 - \frac{(\Sigma M_i\bar{y}_i)^2}{n}}{n(n-1)}\left(1 - \frac{n}{N}\right)}$$

Now, assume that the 426 woodlots of the previous example have a total area of 26,412 acres. Then, because the secondary units are one-fifth acre in size, the total number of secondaries in the population is $M = 132{,}060$. With the same sample data the unbiased estimate of the population mean per unit would be

$$\bar{\bar{y}} = \frac{426}{4(132{,}060)}(888{,}100) = 716.21 \text{ board feet per plot.}$$

The standard error is

$$s_{\bar{\bar{y}}} = \frac{426}{132{,}060}\sqrt{\frac{(374{,}000^2 + \ldots + 267{,}000^2) - \frac{(888{,}100)^2}{4}}{4(3)}} \quad \begin{pmatrix}\text{fpc}\\ \text{ignored}\end{pmatrix}$$

$$= 0.003226\sqrt{4{,}207{,}253{,}958}$$

$$= 209.25 \text{ board feet.}$$

The standard error of the unbiased estimate (209.25) as compared to that of the biased estimate (57.82) shows why the latter is often preferred. But, if the size of all primaries is known, the bias of the biased estimator can be reduced and the precision of the unbiased estimator increased by grouping similar sized primaries and using these estimating procedures in conjunction with stratified sampling.

Systematic Sampling

As the name implies, and as most foresters know, the units included in a systematic sample are selected not at random but according to a pre-specified pattern. Usually the only element of randomization is in the selection of the starting point of the pattern, and even that is often ignored. The most common pattern is a grid having the sample units in equally spaced rows with a constant distance between units within rows.

To the disdain of some statisticians, the vast majority of forest surveys have been made by some form of systematic sampling. There are two reasons: (1) the location of sample units in the field is often easier and cheaper, and (2) there is a feeling that a sample deliberately spread over the entire population will be more representative than a random sample.

Statisticians usually will not argue against the first reason. They are less willing to accept the second. They admit the possibility, sometimes even the probability, that a systematic sample will give a more precise estimate of the true population mean (i.e., be more representative) than would a random sample of the same size. They point out, however, that estimation of the sampling error of a systematic survey requires more knowledge about the population than is usually available, with the result that the sampler can seldom be sure just how precise his estimate is. The common procedure is to use random sampling formulae to compute the errors of a systematic survey. Depending on the degree and the way in which the population falls into patterns, the precision may be either much lower or much higher than that suggested by the random formulae. If there is no definite pattern in the unit values in the population, the random formulae may give a fair indication of the sampling precision. The difficulty is in knowing which condition applies to a particular sample.

The well-known procedure of superimposing two or more systematic grids, each with randomly located starting points, does provide some of the advantages of systematic sampling along with a valid estimate of the sampling error. In this procedure each grid becomes, in effect, a single observation and the error is estimated from the variability among grids. Locating plots in the field becomes more difficult as the number of grids increases, however, and it would seem as though the advantage of representativeness could be obtained more easily and efficiently by stratified sampling with small blocks serving as strata.

Despite the known hazards, foresters are not likely to give up systematic sampling. They will usually take the precaution of running the lines of plots at right angles rather than parallel to ridges and streams. In most cases, sampling errors will be computed by formulae appropriate to random sampling. Experience suggests that a few of these surveys will be very misleading, but that most of them will give estimates having precision as good as or slightly better than that shown by the random sampling formulae. Some statisticians will continue to bemoan the practice and a few of them will keep searching for a workable general solution to the problem of error estimates (though at least one very eminent statistician doubts that a workable solution exists).

SAMPLING METHODS FOR DISCRETE VARIABLES

Simple Random Sampling—Classification Data

Assume that from a large batch of seed 50 have been selected at random in order to estimate the proportion (p) that are sound.

Assume also that cutting or hammering discloses that 39 of the 50 seeds were sound. Then our estimate ($\bar{p}$) of the proportion that is sound is

$$\begin{aligned}\bar{p} &= \frac{\text{Number having the specified attribute}}{\text{Number observed}}\\ &= \frac{39}{50}\\ &= 0.78\end{aligned}$$

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

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

where: n = number of units observed.

In this example N is extremely large relative to n, and so the finite-population correction could be ignored

$$\begin{aligned}s_{\bar{p}} &= \sqrt{\frac{(0.78)(1-0.78)}{(50-1)}}\\ &= 0.05918\end{aligned}$$

Confidence limits.—For certain sample sizes (among them, $n = 50$), confidence limits can be obtained from table 3, page 87. In this example we found that in a sample of $n = 50$ seeds, 39 were sound. The estimated proportion sound 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 $n = 1{,}000$ seeds, 78 percent were sound. This is equivalent to saying that 22 percent were not sound, and the table shows that for $n = 1{,}000$ the 95-percent confidence interval for an observed fraction of 0.22 is 0.19 to 0.25. If the true population proportion of unsound seed is within the limits 0.19 and 0.25, the population proportion of sound seed must be within the limits 0.75 and 0.81.

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

$$\bar{p} \pm \left[2s_{\bar{p}} + \frac{1}{2n}\right]$$

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,

$$\bar{p} = \frac{70}{250} = 0.28$$

And,

$$s_{\bar{p}} = \sqrt{\frac{(0.28)(0.72)}{249}} \quad \text{(ignoring the finite-population correction)}$$
$$= 0.02845$$

Then, the 95-percent confidence interval

$$= 0.28 \pm \left[2(0.02845) + \frac{1}{2(250)}\right]$$
$$= 0.28 \pm 0.059$$
$$= 0.221 \text{ to } 0.339$$

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

The above equation gives what is known as the normal approximation to the confidence limits. As noted, it can be used for large samples. What qualifies as a large sample depends on the proportion of items having the specified characteristic. As a rough guide, the normal approximation will be good if the common 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.

$|P - 0.5|$ = the absolute value (i.e., algebraic sign ignored) of the departure of P from 0.5.

Thus, if our estimate of P is 0.20 then $|P - 0.5|$ is equal to 0.3 and, if we are to use the normal approximation, the log of our sample size should be greater than

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

Or n must be over 251 ($2.4 = \log 251$).

Sample size.—Table 3 may also be used as a guide to the number of units that should be observed in a simple random sample 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 attribute 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 having $\bar{p} = 0.4$ the confidence limits would be 0.23 and 0.60. Since the upper limit is not within 0.15 of $\bar{p} = 0.4$, a sample of size 30 would not give the necessary precision. A sample of $n = 50$ would give limits of 0.27 and 0.55. As each of these is within 0.15 or $\bar{p} = 0.4$, 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

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

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

where: E = The precision with which P is to be estimated.
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 would find,

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

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.

How to select a seed at random.—If we were trying to estimate the proportion of trees in a stand having a certain disease, it would be difficult to select the individual trees at random and then locate them in the field for observation. In some populations, however, 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 together towards the top of the pile. The sample could be taken with a small scoop or a seed probe which picks up approximately the number of seed to be examined. As a precaution, most seed samplers will use a scoop that selects only a fraction of the desired number of seeds and will take samples from several places in the pile and combine them.

Cluster Sampling for Attributes

In attribute sampling the cost of selecting and locating an individual is usually very high relative to the cost of determining whether or not the individual has a certain characteristic. Because of this, some form of cluster sampling is usually preferred over simple random sampling. In cluster sampling, a group of individuals becomes the unit of observation, and the unit value is the proportion of the individuals in the group having the specified attribute.

In estimating the survival percent of a plantation it would be 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 would obviously be inefficient to ignore all of the trees that must be passed to get to the one selected. Instead, we would probably make survival counts in a number of randomly selected rows and (assuming the same number of trees were planted in each row) average these to estimate the survival percent. This is a form of cluster sampling, the cluster being a row of planted trees.

The germination percent of a batch of seed might also be estimated by cluster sampling. Here the advantage of clusters comes not in the selection of individuals 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., overwatering 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 8 separate germination dishes. After 30 days, the following germination percentages are recorded:

Dish No.	1	2	3	4	5	6	7	8	Total
Germination (pct.)	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

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

The variance of p would be computed by

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

Whence the standard error of $\bar{p}$ can be obtained as

$$s_{\bar{p}} = \sqrt{\frac{s_p^2}{n}\left(1 - \frac{n}{N}\right)}$$
$$= \sqrt{\frac{14.5714}{8}} = 1.35 \text{ (ignoring the finite-population correction)}$$

Note that, in cluster sampling, 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 Student's t

$$\text{95-percent confidence interval} = \bar{p} \pm t\ (s_{\bar{p}})$$

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

$$\begin{aligned} 83.0 \pm (2.365)\ (1.35) &= 83.0 \pm 3.19 \\ &= 79.8 \text{ to } 86.2 \end{aligned}$$

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. The common transformation is arcsin $\sqrt{\text{percent}}$. Table 4, page 89, makes it easy to transform the observed percentages. For the data in the previous example, the transformed values would be

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

The mean of the transformed values is

$$\frac{526.0}{8} = 65.75$$

The variance of these values is

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

And the standard error of the mean transformed value is

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

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

$$\begin{aligned} CI = 65.75 \pm (2.365)\ (1.009) &= 65.75 \pm 2.39 \\ &= 63.36 \text{ to } 68.14 \end{aligned}$$

Referring to the table again, we see that the mean of 65.75 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.

With seed stored in a number of canisters of 100 pounds each, we might use two-stage sampling, the canisters being primary sampling units and clusters of 100 seeds being secondaries. If the canisters differed in volume, we might sample canisters with probability proportional to size.

Cluster Sampling for Attributes—Unequal-Sized Clusters

Frequently when sampling for attributes, we find it convenient to let a plot be the sampling unit. On each plot we will 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 probably would not want to average the proportions for all plots because that would give the same weight to plots with few individuals as to those with many.

In such situations, we might use the ratio-of-means estimator. Suppose that 2,4,5-T 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 one-tenth-acre plots.

Plot	*No. of trees* (x_i)	*No. of 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	1,351	960
Mean	67.55	48.0

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

$$\bar{p} = \frac{\bar{y}}{\bar{x}} = \frac{48.0}{67.55} = 0.7106$$

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

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

Where: $s_y{}^2$ = Variance of individual y values.
$s_x{}^2$ = Variance of individual x values.
s_{yx} = Covariance of y and x.
n = Number of plots observed.

In this example

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

$$s_x{}^2 = \frac{(15^2 + 42^2 + \ldots + 84^2) - \dfrac{1{,}351^2}{20}}{19} = 1{,}542.4711$$

$$s_{yx} = \frac{(11)(15) + (32)(42) + \ldots + (59)(84) - \dfrac{(960)(1{,}351)}{20}}{19}$$

$$= 1{,}132.6316$$

With these values (but ignoring the fpc),

$$s_{\bar{p}} = \sqrt{\frac{1}{(67.55)^2}\left[\frac{892.6316 + (0.7106)^2\,(1{,}542.4711) - 2(0.7106)\,(1{,}132.6316)}{20}\right]}$$

$$= 0.026$$

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 one-tenth-milacre 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.) that are difficult to work with. If count variables cannot be avoided, the amateur 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 lärvae of a certain insect in the litter of a forest tract, one-foot-square litter samples were taken at 600 randomly selected points. 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 were

Count	=	0	1	2	3	4	5	6	Total
Number of plots	=	256	224	92	21	4	1	2	600

The counts are very close to a Poisson distribution (see page 6). To permit the applications of normal distribution methods, the units were redefined. The new units were to consist of 15 of the original units selected at random from the 600. There were to be a total of 40 of the new units, and unit values were to be the total larvae count for the 15 selected observations. The values for the 40 redefined units were

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

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

With correction for finite population ignored, the standard error of the mean ($s_{\bar{y}}$) is

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

The new units have a total area of 15 square feet; hence to estimate the mean number of larvae per acre the mean per unit must be multiplied by

$$\frac{43{,}560}{15} = 2{,}904$$

Thus, the mean per acre is

$$(2{,}904)\,(12.6) = 36{,}590.4$$

The standard error of the mean per acre is

$$(2{,}904)\,(0.47) = 1{,}364.88$$

As an approximation we can say that unless a 1-in-20 chance has occurred in sampling, the mean count per acre is within the limits

$$36{,}590.4 \pm 2(1{,}364.88)$$

or,

$$33{,}860 \text{ to } 39{,}320$$

SOME OTHER ASPECTS OF SAMPLING

Size and Shape of Sampling Units

The size and shape of the sampling unit may profoundly affect the cost of the survey, its precision, or both. No attempt will be made here to offer an exhaustive study, but an example may illustrate the problem and a general approach to its solution.

Consider a preharvest inventory in a nursery containing 1,000 beds of slash pine, each bed 500 feet long and 4 feet wide. Conventional practice in this nursery has been to sample the beds by observing the total number of plantable seedlings in a 1- by 4-foot sampling frame laid crosswise at five randomly chosen locations in each bed. The process is laborious and time consuming, totaling 5,000 observations, or nearly a mile of bed. The nurseryman would like to know if a frame 6 inches wide would be better than the conventional 12-inch frame.

One practical way to judge among sampling units is to compare the total cost of surveys made with each unit, with the restriction that both methods shall afford equal precision. For example,[3] if the cost per observation with the 6-inch frame is d_1, then for n_1 observations the cost of the survey (exclusive of overhead costs, which are assumed to be the same for both size units) is

$$c_1 = n_1 d_1$$

Similarly, for the 12-inch frame, we can say

$$c_2 = n_2 d_2$$

[3] For illustrative purposes the nursery survey will be treated as a simple random sample, though the specification of a set number of plots in each bed makes it a stratified design.

Then the cost of the 6-inch frame relative to the cost of the 12-inch frame is

$$\frac{c_1}{c_2} = \frac{n_1 d_1}{n_2 d_2}$$

If estimates of population variance s_1^2 and s_2^2 are available, variance of the population totals (ignoring the fpc) may be written

$$s_{T_1}{}^2 = N_1{}^2 \left(\frac{s_1{}^2}{n_1}\right)$$

and

$$s_{T_2}{}^2 = N_2{}^2 \left(\frac{s_2{}^2}{n_2}\right)$$

where: N_1 and N_2 = Number of units of each size in the population. Now if the two methods are to give equal precision for the estimate of total production,

$$s_{T_1}{}^2 = s_{T_2}{}^2$$

or,

$$N_1{}^2 \left(\frac{s_1{}^2}{n_1}\right) = N_2{}^2 \left(\frac{s_2{}^2}{n_2}\right)$$

and, solving for n_2

$$n_2 = \left(\frac{N_2{}^2 s_2{}^2}{N_1{}^2 s_1{}^2}\right) n_1$$

This last quantity may be simplified by remembering that the total number of 6-inch units (N_1) is twice the total number of 12-inch units (N_2); hence,

$$\begin{aligned} n_2 &= \left(\frac{N_2{}^2 s_2{}^2}{(2N_2)^2 s_1{}^2}\right) n_1 \\ &= \frac{n_1 s_2{}^2}{4 s_1{}^2} \end{aligned}$$

If we substitute this value of n_2 in the relative cost formula given above

$$\begin{aligned} \frac{c_1}{c_2} = \frac{n_1 d_1}{n_2 d_2} &= \frac{n_1 d_1}{\left(\frac{n_1 s_2{}^2}{4 s_1{}^2}\right) d_2} \\ &= \frac{4 s_1{}^2 d_1}{s_2{}^2 d_2} \end{aligned}$$

In this example, a special study showed s_1^2 and s_2^2 to be 134.1 and 416.0 respectively, and the average times for locating the frame and making the count for each size of frame were found to be

$d_1 = 94.36$ and $d_2 = 129.00$. Substituting these values in the equation for relative cost,

$$\frac{c_1}{c_2} = \frac{4(134.1)(94.36)}{(416.0)(129.00)}$$
$$= 0.943$$

This result indicates that the 6-inch frame is slightly more efficient than the 12-inch frame.

In more general terms the cost of method 1 relative to the cost of method 2 for a specified sampling error would be

$$\frac{c_1}{c_2} = \frac{N_1^2 s_1^2 d_1}{N_2^2 s_2^2 d_2}$$

The same result is obtained by thinking in terms of the relative efficiency of the alternative procedures. As a measure of efficiency, statisticians commonly use the reciprocal of the product of the cost per unit and the squared coefficient of variation for the given sample unit. If the coefficient of variation is symbolized by C and the cost by d, the efficiency (U) is given by

$$U = \frac{1}{(d)(C)^2}$$

The relative efficiency of two alternatives would then be

$$\frac{U_2}{U_1} = \frac{(d_1)(C_1)^2}{(d_2)(C_2)^2} \text{ or, } \frac{U_1}{U_2} = \frac{(d_2)(C_2)^2}{(d_1)(C_1)^2}$$

In the previous example we had

$$d_1 = 94.36 \qquad s_1^2 = 134.1$$
$$d_2 = 129.00 \qquad s_2^2 = 416.0$$

For the 6-inch frame the squared coefficient of variation is

$$(C_1)^2 = \frac{s_1^2}{\bar{x}_1^2}$$

For the 12-inch frame the squared coefficient of variation would be

$$(C_2)^2 = \frac{s_2^2}{\bar{x}_2^2}$$

The mean per unit for the 12-inch frame $(\bar{x}_2)$ should be twice the mean per unit for the 6-inch frame, so that we can write

$$(C_2)^2 = \frac{s_2^2}{(2\bar{x}_1)^2} = \frac{s_2^2}{4\bar{x}_1^2}$$

Then the efficiency of the 12-inch frame relative to that of the 6-inch frame is

$$\frac{U_2}{U_1} = \frac{94.36(134.1/\bar{x}_1^2)}{129.00(416.0/4\bar{x}_1^2)} = \frac{4(94.36)(134.1)}{(129.00)(416.0)}$$
$$= 0.943$$

As before, the 6-inch frame appears more efficient than the 12-inch frame.

Estimating Changes

Changes that have taken place in the characteristics of a forest population are often of as much interest as their present status. Periodic change in stand volume is, for example, a major concern of foresters.

Estimating such changes usually requires sampling at the beginning and end of the period. The difference or some function of the difference between the two estimates is the estimated change. Ordinarily the same sampling method will be used each time, but that is not absolutely necessary.

Temporary or permanent plots.—Estimating change by sampling at two different times always raises the question of temporary or permanent sample plots. That is, should an entirely new set of units be randomly selected for observation at each time, or should the same units be observed at both times? A third alternative is to have some temporary and some permanent plots in a double sampling system: a large sample of temporary plots with a subsample of permanent plots.

The choice between temporary and permanent plots depends heavily on the degree of correlation that can be expected between the initial and final plot values. If a high positive correlation is expected, permanent plots should give the better precision. If the correlation is likely to be low or negative, temporary plots might be better. If the period is relatively short and if cutting or heavy mortality is unlikely, the correlation probably will be large and positive, favoring the use of permanent plots. Where large volume changes are likely to occur because of cutting, heavy mortality, or a very long time interval, the correlation will be small or even negative, favoring the use of temporary plots.

If there is enough information on cost and variability, the advantage of permanent plots with simple random sampling can be weighed by computing the relative cost (R_c) of obtaining a given precision by the two methods.

$$R_c = \frac{2C_t(s_1^2 + s_2^2)}{C_p(s_1^2 + s_2^2 - 2s_{12})}$$

where: C_t = Cost of locating and making a single measurement on a temporary plot.

C_p = Total cost of locating, measuring, monumenting, relocating, and remeasuring a permanent plot.

s_1^2 = Variance among individual plots at the time of the first measurement.

s_2^2 = Variance among individual plots at the time of the second measurement.

s_{12} = Covariance between the first and second measurements on individual plots.

If R_c is greater than 1, permanent plots should be used. If R_c is less than 1, temporary plots will probably be better. Where remeasurements will be made several times, the average cost per permanent plot will be reduced, swinging the ratio more favorably towards permanent plots.

Plot monumentation.—The question of kind and degree of plot monumentation has been hotly debated among the users of permanent plots. Where any form of stand treatment is likely to take place between measurements, it is generally conceded that the plot location and form of monumentation should not be discernible to those who make the stand treatments. It is very difficult, if not humanly impossible, to avoid treating plot areas differently from nonplot areas. At the same time, if the monuments are too cleverly concealed, relocation costs will be increased and some plots may not be found at all. Because the difficulty of plot relocation is likely to be related to stand conditions that are in turn related to growth, failure to relocate plots could slightly bias the estimates.

Sampling errors.—If the mean per unit at the time of the first measurement is $\bar{y}_1$, and the mean per unit at the time of the second measurement is $\bar{y}_2$, the estimated periodic change per unit is $(\bar{y}_2 - \bar{y}_1)$.

With temporary plots, the standard error of the estimated change would be

$$s_{(\bar{y}_2 - \bar{y}_1)} = \sqrt{s_{\bar{y}_1}{}^2 + s_{\bar{y}_2}{}^2}$$

where $s_{\bar{y}_1}{}^2$ and $s_{\bar{y}_2}{}^2$ are the squared standard errors of the mean at the time of the first and second measurements. The method of computing $s_{\bar{y}_1}{}^2$ and $s_{\bar{y}_2}{}^2$ would be that appropriate to the particular sampling method used.

With permanent plots, the easiest procedure for computing the standard error is to work with the individual differences. Thus, if y_{1i} stands for the first measurement of the i^{th} permanent plot and y_{2i} stands for the second measurement on that plot, then $d_i = (y_{2i} - y_{1i})$. The standard error of the mean difference is computed from the d_i values with the formula appropriate for the particular sampling method.

Examples.—The above computations will be illustrated for a simple random sample.

Temporary Plots

Initial observations: $n = 8$

$$y_{1i} = 12, 24, 27, 14, 16, 10, 21, 30$$

$$\sum_{i=1}^{8} y_{1i} = 154 \qquad \bar{y}_1 = 19.25$$

$$s_{y_1}{}^2 = 53.9286 \qquad s_{\bar{y}_1}{}^2 = \frac{s_{y_1}{}^2}{n} = 6.74$$

Final observations: $n = 8$

$$y_{2j} = 27, 18, 22, 33, 14, 26, 16, 24$$

$$\sum_{j=1}^{8} y_{2j} = 180 \qquad \bar{y}_2 = 22.50$$

$$s_{y_2}^2 = 40.0000 \qquad s_{\bar{y}_2}^2 = \frac{s_{y_2}^2}{n} = 5.00$$

Then the estimated mean difference is

$$(\bar{y}_2 - \bar{y}_1) = (22.50 - 19.25) = 3.25$$

The standard error of the mean difference is

$$\begin{aligned} s_{(\bar{y}_2 - \bar{y}_1)} &= \sqrt{6.74 + 5.00} \\ &= 3.43 \end{aligned}$$

Permanent Plots

	Permanent Plot No.									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Sum*	*Mean*
Initial observations (y_{1i})...	24	14	16	27	10	30	12	21	154	19.25
Final observations (y_{2i})......	26	18	22	27	14	33	16	24	180	22.50
Differences ($d_i = y_{2i} - y_{1i}$)	2	4	6	0	4	3	4	3	26	3.25

The estimated mean difference is

$$(\bar{y}_2 - \bar{y}_1) = \bar{d} = 3.25$$

The standard error of the mean difference is calculated from the d_i values with the formula for a simple random sample.

$$\begin{aligned} s_d^2 &= \frac{\sum_{i=1}^{8} d_i^2 - \frac{\left(\sum_{i=1}^{8} d_i\right)^2}{n}}{(n-1)} \\ &= \frac{(2^2 + 4^2 + \ldots + 3^2) - \frac{26^2}{8}}{7} \\ &= 3.0714 \end{aligned}$$

$$s_{\bar{d}} = \sqrt{\frac{s_d^2}{n}} = 0.62$$

Design of Sample Surveys

It has been the purpose of this handbook to treat only one segment of the design of sample surveys, that of the sampling method and associated computational procedures. These are the aspects of sampling that seem to be most troublesome to foresters. But several other phases of survey design also deserve attention. Some of the points that should be considered in planning a survey are summarized here.

The objective must be stated.—Specifically, identify the parameter to be estimated and the precision desired. An example of a lucid objective might be: "To estimate the number of plantable slash pine seedlings at the Riedsville Nursery. The estimate should be within 1 percent of the true number, with 95-percent confidence." Vague statements ("To study the results of spraying..." "To estimate the effectiveness of...") can and do result in an appalling waste of survey efforts.

The population should be defined.—What are the units constituting the population? What are the unit values? What units are excluded from the population? Careful, accurate answers to these questions will forestall numerous difficulties at later stages. A generality worth repeating is that sampling design will be simplified if the specifications for the units used to define the population are identical with those used in the sample. Even at that, the definition and specification may be difficult. It may be easy to define a tree or a plot, but if a survey is to be made of farmers, pulpwood contractors, or seed orchards, the unit may be very hard to define. An attempt should be made to foresee the difficulties that might arise in classifying a unit as in or out of the population; the borderline instances will be a constant source of trouble to enumerators and analysts.

The data to be collected should be specified.—Special attention must be paid to getting all the data necessary to the objective. It is a moot question how far one should go in taking supplementary data that is not pertinent to the main objective. Frequently cooperators and reviewers, sensing an opportunity to obtain information on some pet project, will request that additional observations be made "while you're there." Such requests must be carefully reviewed. "Free" information is not cheap if it is never used or has an adverse effect on the main objective of the survey.

Measurement techniques must be prescribed.—The measurement procedures should be stated unambiguously. The detail needed will vary with the complexity of the measurements and the experience of the personnel, but in general it is better to be annoyingly specific than trustingly vague. Terms such as *merchantable top, overstory, undesirable, stocked, board-foot volume,* and *plantable* should be precisely defined.

The need for training and preliminary practice should be considered, And proficiency tests are not unwarranted—even for the old hands who may have forgotten some of their earlier training or developed bad habits.

The sampling units must be defined.—Again, the totality of sampling units, however distributed, must comprise the population. If the unit is obvious, e.g., a sawmill, no particular trouble need arise. But if a variety of units are possible, a search of literature will frequently uncover some profitable experience; if not, a study of the optimum size and shape of sampling unit may be required.

The sampling method must be described.—This handbook outlines a number of methods that have been found useful in forestry. Thought, experience, and a review of literature will help in deciding which method is most appropriate for a particular situation. The method of selecting the sample units should be carefully stated, and so should the procedure of locating the units in the field. Saying that a two-stage design will be used with primaries and secondaries selected at random is not enough. How will randomization be accomplished? And how will the unit be located in the field? The possibilities of and antidotes for bias in locating units deserve some thought. Timber cruisers will, for example, tend to veer away from dense brush and openings when locating plots by hand compass and pacing. House-to-house interviewers have been known to neglect top-floor apartments and homes with barking dogs.

At this stage it is also well to think out the procedures to be used for estimating the parameters and sampling errors. Collecting data and then asking someone how to use it is a good way to lose friends and waste survey money.

The sample size must be prescribed.—Once the desired precision, choice of sampling unit, and method of sampling have been stated it is time to think of the size of sample. The sample should be just large enough to give the specified precision, and no larger. If the requisite information on costs and variances is available, this decision should be made prior to the start of field work. In the absence of such information, a preliminary survey may be necessary.

Possible problems of data should be considered.—If the preceding steps are meticulously followed, problems arising at the data-collection stage are usually those of organization and personnel. The greatest single stumbling block is the common failure of supervisors to continue training and checking field crews or to provide for editing of field forms. Some organizations find it worthwhile to make punched-card sorts to check for recording mistakes such as trees that are 3 inches in d.b.h. and have 14 logs (instead of a 14-inch tree with 3 logs).

Data processing should be planned.—In most cases, procedures for computation and analysis are fixed by the choice of sampling methods. In organizing the computing, there may be some extraordinary considerations that merit early attention. If the volume of data is small, computing may be readily absorbed in the daily routine. If the volume is large, special staffing and special equipment may be desirable. If, for example, the analysis is to be on electronic computers, it would be advisable to become familiar with the special requirements necessary to electronic computing, such as data format for keypunching, availability of programs, and cost of programming.

REFERENCES FOR ADDITIONAL READING

Cochran, W. G.
1953. Sampling techniques. 330 pp., illus. Wiley, New York.

Deming, W. E.
1950. Some theory of sampling. 602 pp., illus. Wiley, New York.

Dixon, W. J., and Massey, F. J., Jr.
1957. Introduction to statistical analysis. Ed. 2, 488 pp., illus. McGraw-Hill, New York.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G.
1953. Sample survey methods and theory. Vol. I, 638 pp., illus. Wiley, New York.

Hendricks, W. A.
1956. The mathematical theory of sampling. 364 pp., illus. Scarecrow Press, New Brunswick, N. J.

Schumacher, F. X., and Chapman, R. A.
1942. Sampling methods in forestry and range management. Duke Univ. School Forestry Bul. 7, 213 pp., illus.

Snedecor, G. W.
1956. Statistical methods. Ed. 5, 534 pp., illus. Iowa State Univ. Press, Ames, Ia.

Sukhatme, P. V.
1954. Sampling theory of surveys, with applications. 491 pp., illus. Iowa State Univ. Press, Ames, Ia.

Yates, Frank.
1960. Sampling methods for censuses and surveys. Ed. 2, 440 pp., illus. Hafner, New York.

PRACTICE PROBLEMS IN SUBSCRIPT AND SUMMATION NOTATION

Values of the Variable x_{ij}

		j Classification (j=1, . . . , 10)										i Classification subtotals
		1	2	3	4	5	6	7	8	9	10	
i Classification i = (1, . . . , 7)	1	6	4	2	0	4	3	5	9	6	8	47
	2	4	8	4	2	1	1	1	6	2	1	30
	3	2	3	2	8	4	8	2	1	1	2	33
	4	1	0	3	2	0	0	0	2	4	8	20
	5	0	2	6	7	1	8	3	5	4	4	40
	6	3	7	5	3	5	2	4	3	2	6	40
	7	2	1	7	2	6	1	1	6	4	3	33
j Classification subtotals		18	25	29	24	21	23	16	32	23	32	243

Examples:

$x_{1,3} = 2 \qquad x_{7,5} = 6 \qquad x_{3,7} = 2 \qquad x_{4,7} = 0$

$$\sum_{i=1}^{7} \sum_{j=1}^{10} x_{ij} = (x_{1,1} + x_{1,2} + \ldots + x_{1,10} + x_{2,1} + x_{2,2} + \ldots + x_{7,10})$$

$$= (6 + 4 + \ldots + 8 + 4 + 8 + \ldots + 3) = 243$$

$$\sum_{i=2}^{3} \sum_{j=1}^{3} x_{ij} = (x_{2,1} + x_{2,2} + x_{2,3} + x_{3,1} + x_{3,2} + x_{3,3})$$

$$= (4 + 8 + 4 + 2 + 3 + 2) = 23$$

$$\sum_{i=1}^{2} \sum_{j=3}^{4} x_{ij}^2 = (x_{1,3}^2 + x_{1,4}^2 + x_{2,3}^2 + x_{2,4}^2)$$

$$= (2^2 + 0^2 + 4^2 + 2^2) = 24$$

$$\sum_{i=2}^{3} \left(\sum_{j=3}^{4} x_{ij}\right)^2 = (x_{2,3} + x_{2,4})^2 + (x_{3,3} + x_{3,4})^2$$

$$= (4 + 2)^2 + (2 + 8)^2 = 136$$

$$\left(\sum_{i=5}^{6}\sum_{j=8}^{9} x_{ij}\right)^2 = (x_{5,8} + x_{5,9} + x_{6,8} + x_{6,9})^2$$
$$= (5 + 4 + 3 + 2)^2 = 196$$

$$\sum_{j=1}^{10} x_{3j} = (x_{3,1} + x_{3,2} + \ldots + x_{3,10})$$
$$= (2 + 3 + \ldots + 2) = 33$$

$$\sum_{i} x_{i6}^2 = (3^2 + 1^2 + 8^2 + \ldots + 1^2) = 143$$

$$\left(\sum_{i} x_{i3}\right)^2 = 29^2 = 841$$

$$\sum_{i,\ j} x_{ij} = 243$$

$$\sum_{i=1}^{7} x_{i2}x_{i3} = (x_{1,2})(x_{1,3}) + (x_{2,2})(x_{2,3}) + \ldots + (x_{7,2})(x_{7,3})$$
$$= (4)(2) + (8)(4) + \ldots + (1)(7) = 100$$

$$\sum_{j}(x_{5j} - x_{4j}) = (x_{5,1} - x_{4,1}) + (x_{5,2} - x_{4,2}) + \ldots + (x_{5,10} - x_{4,10})$$
$$= \left(\sum_{j} x_{5j} - \sum_{j} x_{4j}\right)$$
$$= (40 - 20) = 20$$

$$\sum_{j}(x_{5j} - x_{4j})^2 = (0 - 1)^2 + (2 - 0)^2 + (6 - 3)^2 + \ldots + (4 - 8)^2$$
$$= 138$$

$$\sum_{j} x_{5j}^2 - \sum_{j} x_{4j}^2 = (0^2 + 2^2 + \ldots + 4^2) - (1^2 + 0^2 + \ldots + 8^2) = 122$$

$$\left(\sum_{j} x_{5j}\right)^2 - \left(\sum_{j} x_{4j}\right)^2 = (40^2 - 20^2)$$
$$= 1{,}200$$

$$\left[\sum_{j}(x_{5j} - x_{4j})\right]^2 = \left[\sum_{j} x_{5j} - \sum_{j} x_{4j}\right]^2$$
$$= [40 - 20]^2 = 400$$

$$\begin{aligned}
\sum_j 3x_{2j} &= 3(x_{2,1}) + 3(x_{2,2}) + \ldots + 3(x_{2,10}) \\
&= 3(x_{2,1} + x_{2,2} + \ldots + x_{2,10}) \\
&= 3\left(\sum_j x_{2j}\right) = 3(30) = 90
\end{aligned}$$

$$\begin{aligned}
\sum_j (x_{4j} - 6) &= (x_{4,1} - 6) + (x_{4,2} - 6) \\
&\qquad + \ldots + (x_{4,10} - 6) \\
&= (x_{4,1} + x_{4,2} + \ldots + x_{4,10}) \\
&\qquad - 6 - 6 - \ldots - 6 \\
&= \left(\sum_j x_{4j}\right) - 10(6) \\
&= (20 - 60) = -40
\end{aligned}$$

TABLE 1.—*Ten thousand randomly assorted digits*

	00–04	05–09	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186	59391	58030	52098	82718	87024	82848	04190	96574	90464	29065
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927	99567	76364	77204	04615	27062	96621	43918	01896	83991	51141
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345	10363	97518	51400	25670	98342	61891	27101	37855	06235	33316
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455	86859	19558	64432	16706	99612	59798	32803	67708	15297	28612
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577	11258	24591	36863	55368	31721	94335	34936	02566	80972	08188
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976	95068	88628	35911	14530	33020	80428	39936	31855	34334	64865
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305	54463	47237	73800	91017	36239	71824	83671	39892	60518	37092
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779	16874	62677	57412	13215	31389	62233	80827	73917	82802	84420
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279	92494	63157	76593	91316	03505	72389	96363	52887	01087	66091
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231	15669	56689	35682	40844	53256	81872	35213	09840	34471	74441
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083	99116	75486	84989	23476	52967	67104	39495	39100	17217	74073
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113	15696	10703	65178	90637	63110	17622	53988	71087	84148	11670
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182	97720	15369	51269	69620	03388	13699	33423	67453	43269	56720
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551	11666	13841	71681	98000	35979	39719	81899	07449	47985	46967
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491	71628	73130	78783	75691	41632	09847	61547	18707	85489	69944
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770	40501	51089	99943	91843	41995	88931	73631	69361	05375	15417
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104	22518	55576	98215	82068	10798	86211	36584	67466	69373	40054
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771	75112	30485	62173	02132	14878	92879	22281	16783	86352	00077
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106	80327	02671	98191	84342	90813	49268	95441	15496	20168	09271
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067	60251	45548	02146	05597	48228	81366	34598	72856	66762	17002
20	75884	12952	84318	95108	72305	64620	91318	89872	45375	85436	57430	82270	10421	05540	43648	75888	66049	21511	47676	33444
21	16777	37116	58550	42958	21460	43910	01175	87894	81378	10620	73528	39559	34434	88596	54086	71693	43132	14414	79949	85193
22	46230	43877	80207	88877	89380	32992	91380	03164	98656	59337	25991	65959	70769	64721	86413	33475	42740	06175	82758	66248
23	42902	66892	46134	01432	94710	23474	20423	60137	60609	13119	78388	16638	09134	59880	63806	48472	39318	35434	24057	74739
24	81007	00333	39693	28039	10154	95425	39220	19774	31782	49037	12477	09965	96657	57994	59439	76330	24596	77515	09577	91871

This table is reproduced, by permission of the author and publishers, from table 1.5.1 of Snedecor's *Statistical Methods* (5th ed.), Iowa State University Press.

TABLE 1.—*Ten thousand randomly assorted digits* (continued)

	00–04	05–09	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99
25	68089	01122	51111	72373	06902	74373	96199	97017	41273	21546	83266	32883	42451	15579	38155	29793	40914	65990	16255	17777
26	20411	67081	89950	16944	93054	87687	96693	87236	77054	33848	76970	80876	10237	39515	79152	74798	39357	09054	73579	92359
27	58212	13160	06468	15718	82627	76999	05999	58680	96739	63700	37074	65198	44785	68624	98336	84481	97610	78735	46703	98265
28	70577	42866	24969	61210	76046	67699	42054	12696	93758	03283	83712	06514	30101	78295	54656	85417	43189	60048	72781	72606
29	94522	74358	71659	62038	79643	79169	44741	05437	39038	13163	20287	56862	69727	94443	64936	08366	27227	05158	50326	59566
30	42626	86819	85651	88678	17401	03252	99547	32404	17918	62880	74261	32592	86538	27041	65172	85532	07571	80609	39285	65340
31	16051	33763	57194	16752	54450	19031	58580	47629	54132	60631	64081	49863	08478	96001	18888	14810	70545	89755	59064	07210
32	08244	27647	33851	44705	94211	46716	11738	55784	95374	72655	05617	75818	47750	67814	29575	10526	66192	44464	27058	40467
33	59497	04392	09419	89964	51211	04894	72882	17805	21896	83864	26793	74951	95466	74307	13330	42664	85515	20632	05497	33625
34	97155	13428	40293	09985	58434	01412	69124	82171	59058	82859	65988	72850	48737	54719	52056	01596	03845	35067	03134	70322
35	98409	66162	95763	47420	20792	61527	20441	39435	11859	41567	27366	42271	44300	73399	21105	03280	73457	43093	05192	48657
36	45476	84882	65109	96597	25930	66790	65706	61203	53634	22557	56760	10909	98147	34736	33863	95256	12731	66598	50771	83665
37	89300	69700	50741	30329	11658	23166	05400	66669	48708	03887	72880	43338	93643	58904	59543	23943	11231	83268	65938	81581
38	50051	95137	91631	66315	91428	12275	24816	68091	71710	33258	77888	38100	03062	58103	47961	83841	25878	23746	55903	44115
39	31753	85178	31310	89642	98364	02306	24617	09609	83942	22716	28440	07819	21580	51459	47971	29882	13990	29226	23608	15873
40	79152	53829	77250	20190	56535	18760	69942	77448	33278	48805	63525	94441	77033	12147	51054	49955	58312	76923	96071	05813
41	44560	38750	83635	56540	64900	42912	13953	79149	18710	68618	47606	93410	16359	89033	89696	47231	64498	31776	05383	39902
42	68328	83378	63369	71381	39564	05615	42451	64559	97501	65747	52669	45030	96279	14709	52372	87832	02735	50803	72744	88208
43	46939	38689	58625	08342	30459	85863	20781	09284	26333	91777	16738	60159	07425	62369	07515	82721	37875	71153	21315	00132
44	83544	86141	15707	96256	23068	13782	08467	89469	93842	55349	59348	11695	45751	15865	74739	05572	32688	20271	65128	14551
45	91621	00881	04900	54224	46177	55309	17852	27491	89415	23466	12900	71775	29845	60774	94924	21810	38636	33717	67598	82521
46	91896	67126	04151	03795	59077	11848	12630	98375	52068	60142	75086	23537	49939	33595	13484	97588	28617	17979	70749	35234
47	55751	62515	21108	80830	02263	29303	37204	96926	30506	09808	99495	51434	29181	09993	38190	42553	68922	52125	91077	40197
48	85156	87689	95493	88842	00664	55017	55539	17771	69448	87530	26075	31671	45386	36583	93459	48599	52022	41330	60651	91321
49	07521	56898	12236	60277	39102	62315	12239	07105	11844	01117	13636	93596	23377	51133	95126	61496	42474	45141	46660	42338

This table is reproduced, by permission of the author and publishers, from table 1.5.1 of Snedecor's *Statistical Methods* (5th ed.), Iowa State University Press.

TABLE 1.—*Ten thousand randomly assorted digits* (continued)

	00–04	05–09	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99
50	64249	63664	39652	40646	97306	31741	07294	84149	46797	82487	32847	31282	03345	89593	69214	70381	78285	20054	91018	16742
51	26538	44249	04050	48174	65570	44072	40192	51153	11397	58212	16916	00041	30236	55023	14253	76582	12092	86533	92426	37655
52	05845	00512	78630	55328	18116	69296	91705	86224	29503	57071	66176	34047	21005	27137	03191	48970	64625	22394	39622	79085
53	74897	68373	67359	51014	33510	83048	17056	72506	82949	54600	46299	13335	12180	16861	38043	59292	62675	63631	37020	78195
54	20872	54570	35017	88132	25730	22626	86723	91691	13191	77212	22847	47839	45385	23289	47526	54098	45683	55849	51575	64689
55	31432	96156	89177	75541	81355	24480	77243	76690	42507	84362	41851	54160	92320	69936	34803	92479	33399	71160	64777	83378
56	66890	61505	01240	00660	05873	13568	76082	79172	57913	93448	28444	59497	91586	95917	68553	28639	06455	34174	11130	91994
57	48194	57790	79970	33106	86904	48119	52503	24130	72824	21627	47520	62378	98855	83174	13088	16561	68559	26679	06238	51254
58	11303	87118	81471	52936	08555	28420	49416	44448	04269	27029	34978	63271	13142	82681	05271	08822	06490	44984	49307	62717
59	54374	57325	16947	45356	78371	10563	97191	53798	12693	27928	37404	80416	69035	92980	49486	74378	75610	74976	70056	15478
60	64852	34421	61046	90849	13966	39810	42699	21753	76192	10508	32400	65482	52099	53676	74648	94148	65095	69597	52771	71551
61	16309	20384	09491	91588	97720	89846	30376	76970	23063	35894	89262	86332	51718	70663	11623	29834	79820	73002	84886	03591
62	42587	37065	24526	72602	57589	98131	37292	05967	26002	51945	86866	09127	98021	03871	27789	58444	44832	36505	40672	30180
63	40177	98590	97161	41682	84533	67588	62036	49967	01990	72308	90814	14833	08759	74645	05046	94056	99094	65091	32663	73040
64	82309	76128	93965	26743	24141	04838	40254	26065	07938	76236	19192	82756	20553	58446	55376	88914	75096	26119	83898	43816
65	79788	68243	59732	04257	27084	14743	17520	95401	55811	76099	77585	52593	56612	95766	10019	29531	73064	20953	53523	58136
66	40538	79000	89559	25026	42274	23489	34502	75508	06059	86682	23757	16364	05096	03192	62386	45389	85332	18877	55710	96459
67	64016	73598	18609	73150	62463	33102	45205	87440	96767	67042	45989	96257	23850	26216	23309	21526	07425	50254	19455	29315
68	49767	12691	17903	93871	99721	79109	09425	26904	07419	76013	92970	94243	07316	41467	64837	52406	25225	51553	31220	14032
69	76974	55108	29795	08404	82684	00497	51126	79935	57450	55671	74346	59596	40088	98176	17896	86900	20249	77753	19099	48885
70	23854	08480	85983	96025	50177	64610	99425	62291	86943	21541	87646	41309	27636	45153	29988	94770	07255	70908	05340	99751
71	68973	70551	25098	78033	98573	79848	31778	29555	61446	23037	50099	71038	45146	06146	55211	99429	43169	66259	97786	59180
72	36444	93600	65350	14971	25325	00427	52073	64280	18847	24768	10127	46900	64984	75348	04115	33624	68774	60013	35515	62556
73	03003	87800	07391	11594	21196	00781	32550	57158	58887	73041	67995	81977	18984	64091	02785	27762	42529	97144	80407	64524
74	17540	26188	36647	78386	04558	61463	57842	90382	77019	24210	26304	80217	84934	82657	69291	35397	98714	35104	08187	48109

This table is reproduced, by permission of the author and publishers, from table 1.5.1 of Snedecor's *Statistical Methods* (5th ed.), Iowa State University Press.

TABLE 1.—*Ten thousand randomly assorted digits* (continued)

	00–04	05–09	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99
75	38916	55809	47982	41968	69760	79422	80154	91486	19180	15100	81994	41070	56642	64091	31229	02595	13513	45148	78722	30144
76	64288	19843	69122	42502	48508	28820	59933	72998	99942	10515	59537	34662	79631	89403	65212	09975	06118	86197	58208	16162
77	86809	51564	38040	39418	49915	19000	58050	16899	79952	57849	51228	10937	62396	81460	47331	91403	95007	06047	16846	64809
78	99800	99566	14742	05028	30033	94889	53381	23656	75787	59223	31089	87995	29577	07828	42272	54016	21950	86192	99046	84864
79	92345	31890	95712	08279	91794	94068	49337	88674	35355	12267	38207	97938	93459	75174	79460	55436	57206	87644	21296	43395
80	90363	65162	32245	82279	79256	80834	06088	99462	56705	06118	88666	31142	09474	89712	63153	62333	42212	06140	42594	43671
81	64437	32242	48431	04835	39070	59702	31508	60935	22390	52246	53365	56134	67582	92557	89520	33452	05134	70628	27612	33738
82	91714	53662	28373	34333	55791	74758	51144	18827	10704	76803	89807	74530	38004	90102	11693	90257	05500	79920	62700	43325
83	20902	17646	31391	31459	33315	03444	55743	74701	58851	27427	18682	81038	85662	90915	91631	22223	91588	80744	07716	12548
84	12217	86007	70371	52281	14510	76094	96579	54853	78339	20839	63571	32579	63942	25371	09234	94592	98475	76884	37635	33608
85	45177	02863	42307	53571	22532	74921	17735	42201	80540	54721	68927	56492	67799	95398	77642	54913	91853	08424	81450	76229
86	28325	90814	08804	52746	47913	54577	47525	77705	95330	21866	56401	63186	39389	88798	31356	89235	97036	32341	33292	73757
87	29019	28776	56116	54791	64604	08815	46049	71186	34650	14994	24333	95603	02359	72942	46287	95382	08452	62862	97869	71775
88	84979	81353	56219	67062	26146	82567	33122	14124	46240	92973	17025	84202	95199	62272	06366	16175	97577	99304	41587	03686
89	50371	26347	48513	63915	11158	25563	91915	18431	92978	11591	02804	08253	52133	20224	68034	50865	57868	22343	55111	03607
90	53422	06825	69711	67950	64716	18003	49581	45378	99878	61130	08298	03879	20995	19850	73090	13191	18963	82244	78479	99121
91	67453	35651	89316	41620	32048	70225	47597	33137	31443	51445	59883	01785	82403	96062	03785	03488	12970	64896	38336	30030
92	07294	85353	74819	23445	68237	07202	99515	62282	53809	26685	46982	06682	62864	91837	74021	89094	39952	64158	79614	78235
93	79544	00302	45338	16015	66613	88968	14595	63836	77716	79596	31121	47266	07661	02051	67599	24471	69843	83696	71402	76287
94	64144	85442	82060	46471	24162	39500	87351	36637	42833	71875	97867	56641	63416	17577	30161	87320	37752	73276	48969	41915
95	90919	11883	58318	00042	52402	28210	34075	33272	00840	73268	57364	86746	08415	14621	49430	22311	15836	72492	49372	44103
96	06670	57353	86275	92276	77591	46924	60839	55437	03183	13191	09559	26263	69511	28064	75999	44540	13337	10918	79846	54809
97	36634	93976	52062	83678	41256	60948	18685	48992	19462	96062	53873	55571	00608	42661	91332	63956	74087	59008	47493	99581
98	75101	72891	85745	67106	26010	62107	60885	37503	55461	71213	35531	19162	86406	05299	77511	24311	57257	22826	77555	05941
99	05112	71222	72654	51583	05228	62056	57390	42746	39272	96659	28229	88629	25695	94932	30721	16197	78742	34974	97528	45447

This table is reproduced, by permission of the author and publishers, from table 1.5.1 of Snedecor's *Statistical Methods* (5th ed.), Iowa State University Press.

TABLE 2.—*The distribution of t*

df	Probability								
	.5	.4	.3	.2	.1	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

This table is abridged from table III of Fisher and Yates' *Statistical Tables for Biological, Agricultural, and Medical Research,* Oliver and Boyd Ltd., Edinburgh. Permission has been given by the authors and publishers.

TABLE 3.—*Confidence intervals for binominal distribution*

Number observed (*f*)	95-percent interval																
	Size of sample, *n*												Fraction observed *f*/*n*	Size of sample			
	10		15		20		30		50		100			250		1000	
0	0	31	0	22	0	17	0	12	0	07	0	4	0.00	0	1	0	0
1	0	45	0	32	0	25	0	17	0	11	0	5	.01	0	4	0	2
2	3	56	2	40	1	31	1	22	0	14	0	7	.02	1	5	1	3
3	7	65	4	48	3	38	2	27	1	17	1	8	.03	1	6	2	4
4	12	74	8	55	6	44	4	31	2	19	1	10	.04	2	7	3	5
5	19	81	12	62	9	49	6	35	3	22	2	11	.05	3	9	4	7
6	26	88	16	68	12	54	8	39	5	24	2	12	.06	3	10	5	8
7	35	93	21	73	15	59	10	43	6	27	3	14	.07	4	11	6	9
8	44	97	27	79	19	64	12	46	7	29	4	15	.08	5	12	6	10
9	55	100	32	84	23	68	15	50	9	31	4	16	.09	6	13	7	11
10	69	100	38	88	27	73	17	53	10	34	5	18	.10	7	14	8	12
11	-------		45	92	32	77	20	56	12	36	5	19	.11	7	16	9	13
12	-------		52	96	36	81	23	60	13	38	6	20	.12	8	17	10	14
13	-------		60	98	41	85	25	63	15	41	7	21	.13	9	18	11	15
14	-------		68	100	46	88	28	66	16	43	8	22	.14	10	19	12	16
15	-------		78	100	51	91	31	69	18	44	9	24	.15	10	20	13	17
16	-------		-------		56	94	34	72	20	46	9	25	.16	11	21	14	18
17	-------		-------		62	97	37	75	21	48	10	26	.17	12	22	15	19
18	-------		-------		69	99	40	77	23	50	11	27	.18	13	23	16	21
19	-------		-------		75	100	44	80	25	53	12	28	.19	14	24	17	22
20	-------		-------		83	100	47	83	27	55	13	29	.20	15	26	18	23
21	-------		-------		-------		50	85	28	57	14	30	.21	16	27	19	24
22	-------		-------		-------		54	88	30	59	14	31	.22	17	28	19	25
23	-------		-------		-------		57	90	32	61	15	32	.23	18	29	20	26
24	-------		-------		-------		61	92	34	63	16	33	.24	19	30	21	27
25	-------		-------		-------		65	94	36	64	17	35	.25	20	31	22	28
26	-------		-------		-------		69	96	37	66	18	36	.26	20	32	23	29
27	-------		-------		-------		73	98	39	68	19	37	.27	21	33	24	30
28	-------		-------		-------		78	99	41	70	19	38	.28	22	34	25	31
29	-------		-------		-------		83	100	43	72	20	39	.29	23	35	26	32
30	-------		-------		-------		88	100	45	73	21	40	.30	24	36	27	33
31	-------		-------		-------		-------		47	75	22	41	.31	25	37	28	34
32	-------		-------		-------		-------		50	77	23	42	.32	26	38	29	35
33	-------		-------		-------		-------		52	79	24	43	.33	27	39	30	36
34	-------		-------		-------		-------		54	80	25	44	.34	28	40	31	37
35	-------		-------		-------		-------		56	82	26	45	.35	29	41	32	38
36	-------		-------		-------		-------		57	84	27	46	.36	30	42	33	39
37	-------		-------		-------		-------		59	85	28	47	.37	31	43	34	40
38	-------		-------		-------		-------		62	87	28	48	.38	32	44	35	41
39	-------		-------		-------		-------		64	88	29	49	.39	33	45	36	42
40	-------		-------		-------		-------		66	90	30	50	.40	34	46	37	43
41	-------		-------		-------		-------		69	91	31	51	.41	35	47	38	44
42	-------		-------		-------		-------		71	93	32	52	.42	36	48	39	45
43	-------		-------		-------		-------		73	94	33	53	.43	37	49	40	46
44	-------		-------		-------		-------		76	95	34	54	.44	38	50	41	47
45	-------		-------		-------		-------		78	97	35	55	.45	39	51	42	48
46	-------		-------		-------		-------		81	98	36	56	.46	40	52	43	49
47	-------		-------		-------		-------		83	99	37	57	.47	41	53	44	50
48	-------		-------		-------		-------		86	100	38	58	.48	42	54	45	51
49	-------		-------		-------		-------		89	100	39	59	.49	43	55	46	52
50	-------		-------		-------		-------		93	100	40	60	.50	44	56	47	53

This table is reproduced, by permission of the author and publishers, from table 1.3.1 of Snedecor's *Statistical Methods* (ed. 5), Iowa-State University Press.

TABLE 3.—*Confidence intervals for binominal distribution* (continued)

Number observed (*f*)	99-percent interval																
	Size of sample, *n*												Fraction observed *f/n*	Size of sample			
	10		15		20		30		50		100			250		1000	
0	0	41	0	30	0	23	0	16	0	10	0	5	0.00	0	2	0	1
1	0	54	0	40	0	32	0	22	0	14	0	7	.01	0	5	0	2
2	1	65	1	49	1	39	0	28	0	17	0	9	.02	1	6	1	3
3	4	74	2	56	2	45	1	32	1	20	0	10	.03	1	7	2	4
4	8	81	5	63	4	51	3	36	1	23	1	12	.04	2	9	3	6
5	13	87	8	69	6	56	4	40	2	26	1	13	.05	2	10	3	7
6	19	92	12	74	8	61	6	44	3	29	2	14	.06	3	11	4	8
7	26	96	16	79	11	66	8	48	4	31	2	16	.07	3	13	5	9
8	35	99	21	84	15	70	10	52	6	33	3	17	.08	4	14	6	10
9	46	100	26	88	18	74	12	55	7	36	3	18	.09	5	15	7	12
10	59	100	31	92	22	78	14	58	8	38	4	19	.10	6	16	8	13
11	-------		37	95	26	82	16	62	10	40	4	20	.11	6	17	9	14
12	-------		44	98	30	85	18	65	11	43	5	21	.12	7	18	9	15
13	-------		51	99	34	89	21	68	12	45	6	23	.13	8	19	10	16
14	-------		60	100	39	92	24	71	14	47	6	24	.14	9	20	11	17
15	-------		70	100	44	94	26	74	15	49	7	26	.15	9	22	12	18
16	-------		-------		49	96	29	76	17	51	8	27	.16	10	23	13	19
17	-------		-------		55	98	32	79	18	53	9	29	.17	11	24	14	20
18	-------		-------		61	99	35	82	20	55	9	30	.18	12	25	15	21
19	-------		-------		68	100	38	84	21	57	10	31	.19	13	26	16	22
20	-------		-------		77	100	42	86	23	59	11	32	.20	14	27	17	23
21	-------		-------		-------		45	88	24	61	12	33	.21	15	28	18	24
22	-------		-------		-------		48	90	26	63	12	34	.22	16	30	19	26
23	-------		-------		-------		52	92	28	65	13	35	.23	17	31	20	27
24	-------		-------		-------		56	94	29	67	14	36	.24	18	32	21	28
25	-------		-------		-------		60	96	31	69	15	38	.25	18	33	22	29
26	-------		-------		-------		64	97	33	71	16	39	.26	19	34	22	30
27	-------		-------		-------		68	99	35	72	16	40	.27	20	35	23	31
28	-------		-------		-------		72	100	37	74	17	41	.28	21	36	24	32
29	-------		-------		-------		78	100	39	76	18	42	.29	22	37	25	33
30	-------		-------		-------		84	100	41	77	19	43	.30	23	38	26	34
31	-------		-------		-------		-------		43	79	20	44	.31	24	39	27	35
32	-------		-------		-------		-------		45	80	21	45	.32	25	40	28	36
33	-------		-------		-------		-------		47	82	21	46	.33	26	41	29	37
34	-------		-------		-------		-------		49	83	22	47	.34	26	42	30	38
35	-------		-------		-------		-------		51	85	23	48	.35	27	43	31	39
36	-------		-------		-------		-------		53	86	24	49	.36	28	44	32	40
37	-------		-------		-------		-------		55	88	25	50	.37	29	45	33	41
38	-------		-------		-------		-------		57	89	26	51	.38	30	46	34	42
39	-------		-------		-------		-------		60	90	27	52	.39	31	47	35	43
40	-------		-------		-------		-------		62	92	28	53	.40	32	48	36	44
41	-------		-------		-------		-------		64	93	29	54	.41	33	50	37	45
42	-------		-------		-------		-------		67	94	29	55	.42	34	51	38	46
43	-------		-------		-------		-------		69	96	30	56	.43	35	52	39	47
44	-------		-------		-------		-------		71	97	31	57	.44	36	53	40	48
45	-------		-------		-------		-------		74	98	32	58	.45	37	54	41	49
46	-------		-------		-------		-------		77	99	33	59	.46	38	55	42	50
47	-------		-------		-------		-------		80	99	34	60	.47	39	55	43	51
48	-------		-------		-------		-------		83	100	35	61	.48	40	56	44	52
49	-------		-------		-------		-------		86	100	36	62	.49	41	57	45	53
50	-------		-------		-------		-------		90	100	37	63	.50	42	58	46	54

TABLE 4.—*Arcsin transformation (angles corresponding to percentages, angle = arcsin $\sqrt{percentage}$)*

%	0	1	2	3	4	5	6	7	8	9
0.0......	0	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	4.09	4.13	4.17	4.21	4.25	4.29	4.33	4.37	4.40
0.6......	4.44	4.48	4.52	4.55	4.59	4.62	4.66	4.69	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	7.04	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	9.81
3........	9.98	10.14	10.31	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	12.39	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........	18.44	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	22.55	22.63	22.71
15........	22.79	22.87	22.95	23.03	23.11	23.19	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	25.48	25.55	25.62	25.70	25.77
19........	25.84	25.92	25.99	26.06	26.13	26.21	26.28	26.35	26.42	26.49
20........	26.56	26.64	26.71	26.78	26.85	26.92	26.99	27.06	27.13	27.20
21........	27.28	27.35	27.42	27.49	27.56	27.63	27.69	27.76	27.83	27.90
22........	27.97	28.04	28.11	28.18	28.25	28.32	28.38	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	33.89	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........	37.47	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	40.46	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

TABLE 4.—*Arcsin transformation (angles corresponding to percentages, angle = arcsin* $\sqrt{\textit{percentage}}$*)* (continued)

%	0	1	2	3	4	5	6	7	8	9
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	44.25	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	45.75	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
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	49.54
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	52.53	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	56.11
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	61.62	61.68	61.75	61.82	61.89	61.96
78	62.03	62.10	62.17	62.24	62.31	62.37	62.44	62.51	62.58	62.65
79	62.72	62.80	62.87	62.94	63.01	63.08	63.15	63.22	63.29	63.36
80	63.44	63.51	63.58	63.65	63.72	63.79	63.87	63.94	64.01	64.08
81	64.16	64.23	64.30	64.38	64.45	64.52	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	66.81	66.89	66.97	67.05	67.13
85	67.21	67.29	67.37	67.45	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	71.56	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	72.84	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	77.61	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	79.69	79.86
97	80.02	80.19	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	82.96	83.20	83.45	83.71	83.98

TABLE 4.—*Arcsin transformation (angles corresponding to percentages, angle = arcsin $\sqrt{percentage}$)* (continued)

%	0	1	2	3	4	5	6	7	8	9
99.0	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	85.31	85.34	85.38	85.41	85.45	85.48	85.52
99.4	85.56	85.60	85.63	85.67	85.71	85.75	85.79	85.83	85.87	85.91
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
100.0	90.00	—	—	—	—	—	—	—	—	—

This table is reproduced, by permission of the author and publishers, from table 11.12.1 of Snedecor's *Statistical Methods* (ed. 5), Iowa State University Press. Permission has also been granted by the original author, Dr. C. I. Bliss, of the Connecticut Agricultural Experiment Station.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1962—642443