A Statistics Primer For Foresters

The science of using data isn't just for researchers.

By Susan G. Stafford

Statisticians have been accused of more than their fair share of impropriety. The poet W.H. Auden quipped, "Thou shalt not sit with statisticians nor commit a social science." Books have been written with the disparaging titles of *How to Lie with Statistics* (Huff 1954) and *How to Use (and Misuse) Statistics* (Kimble 1978). Numerous other examples exist of this implied mistrust of statistics and statisticians.

The notion that we can prove anything by manipulating numbers with statistics is a popular misconception. In fact, the objective of statistics—the science that studies the collection and interpretation of numerical data (Clarke and Cooke 1978)—is to answer a general question on the basis of only specific, limited information. For example, we could use statistics to help assess various root-wrenching **treatments** (**boldfaced terms** are defined in the glossary on page 157). How do the treatments affect survival, growth, and morphology of nursery seedlings? We could also determine how various site preparation treatments affect growth and survival of trees in a reforestation unit.

Because information is rarely, if ever, complete, we must rely on statistics: Is our small-scale observation likely to occur on a large scale? For instance, is what we observe in a **sample** of two-year-old Douglas-fir seedlings likely to occur in the **population** of all two-year-old Douglas-fir seedlings in the Northwest? Or was it only a chance occurrence?

Statistics is only a tool. Like any other tool, it can produce meaningful results when properly used and incor-

Susan G. Stafford is assistant professor and forest biometrician, Department of Forest Science, Oregon State University, Corvallis. Paper 1874, Forest Research Laboratory, Oregon State University.

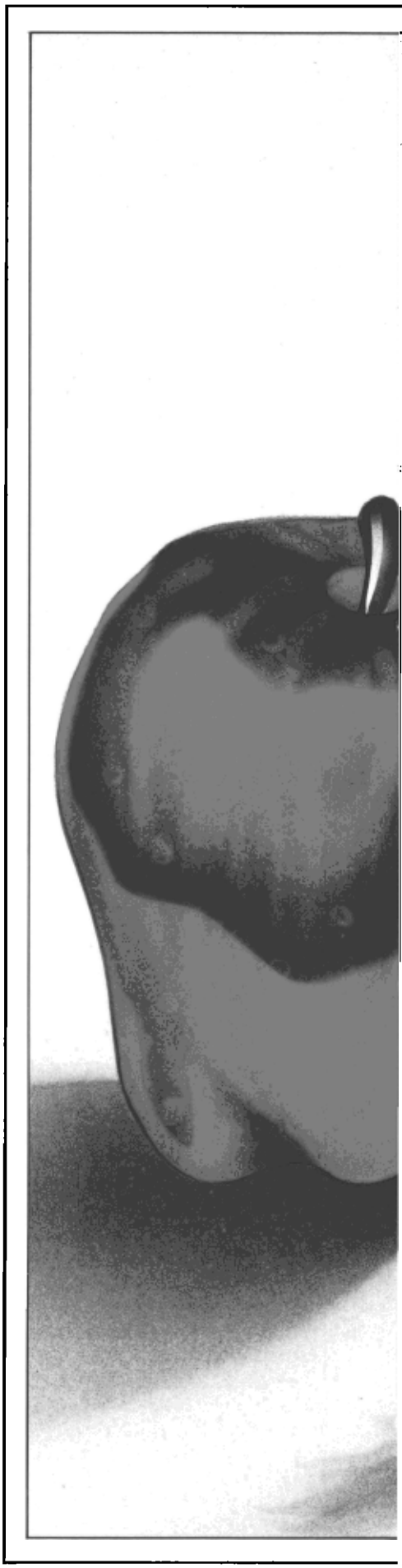
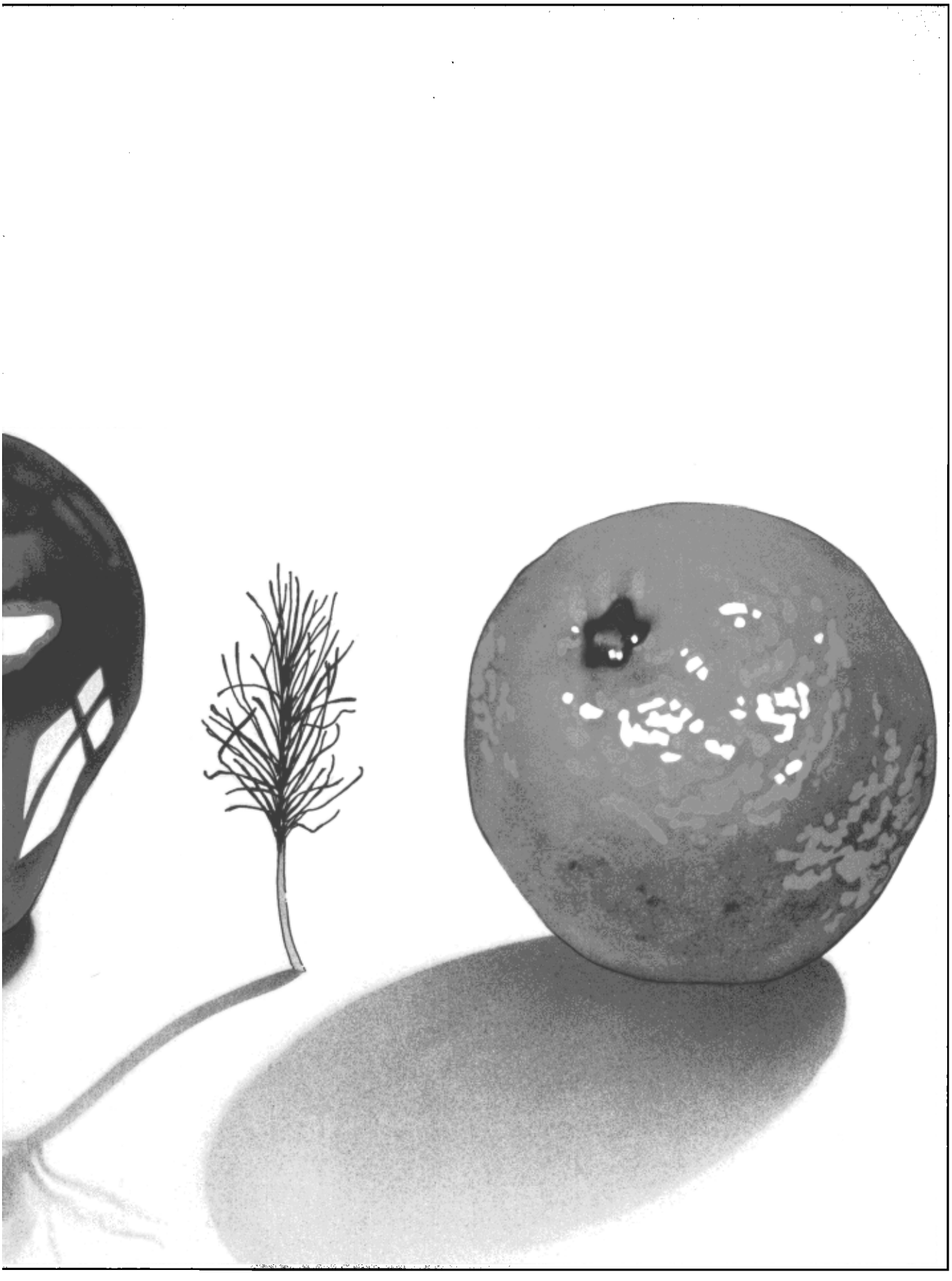


ILLUSTRATION BY BYRON PECK



rect ones when misused. Under the right conditions, however, statistics can be a powerful analytical technique for forest managers, helping them determine the best answers to important questions on the basis of the information given; and, in the process, avoid the costly mistakes made when decisions must be based on limited information. This article addresses what those "right conditions" should be.

Terms and Concepts

Every **experiment** should start with a question to be answered. For example, do different root-wrenching treatments affect seedling growth differently? Or, how do different site-preparation techniques affect seedling survival? Statisticians state these questions as **hypotheses.** To test the validity of a given hypothesis, we draw a sample from a population *(fig. 1a)*, conduct an appropriately designed experiment, and draw inferences based on the data gathered in that experiment and on the experimenter's interpretation of that data. The sample selected should be representative of the population if the inferences drawn are to be correct. Note that an experimenter is not necessarily a researcher, but can be a nursery manager, field forester, harvesting specialist, or certified silviculturist—in short, anyone posing a question for investigation.

We first need to determine the population of interest. For example, in the root-wrenching case, the population of interest might be all conifer seedlings growing in nurseries, all conifer seedlings growing in Pacific Northwest nurseries, or all conifers of a particular species growing in a single nursery. The population can be very large or very small depending upon the question to be answered.

We can add sideboards to the populations and hypotheses, which have the effect of limiting our conclusions. Or we can remove those sideboards, which will make our conclusions more general and give them a wider **scope of inference.** Usually time, money, and materials determine how wide a scope of inference a study can have. For example, we could restrict a root-wrenching study to only a few seedlings from one seedlot grown in one seedbed in one nursery. Although this restriction would be rather extreme, it is acceptable as long as we keep our sideboards in mind at the study's end and do not try to ascribe far-reaching applicability to our narrowly derived results. That is, it would be grossly misleading to apply results from a few seedlings in one seedbed in one nursery to all seedlings in all nurseries.

Once populations and hypotheses are defined, we must determine what to measure and compare. Typically, foresters are interested in the **mean** measurement *(fig. 1b)*—for example, mean height of all two-year-old Douglas-fir seedlings in Pacific Northwest nurseries. The mean is a numerical characteristic describing the population and is

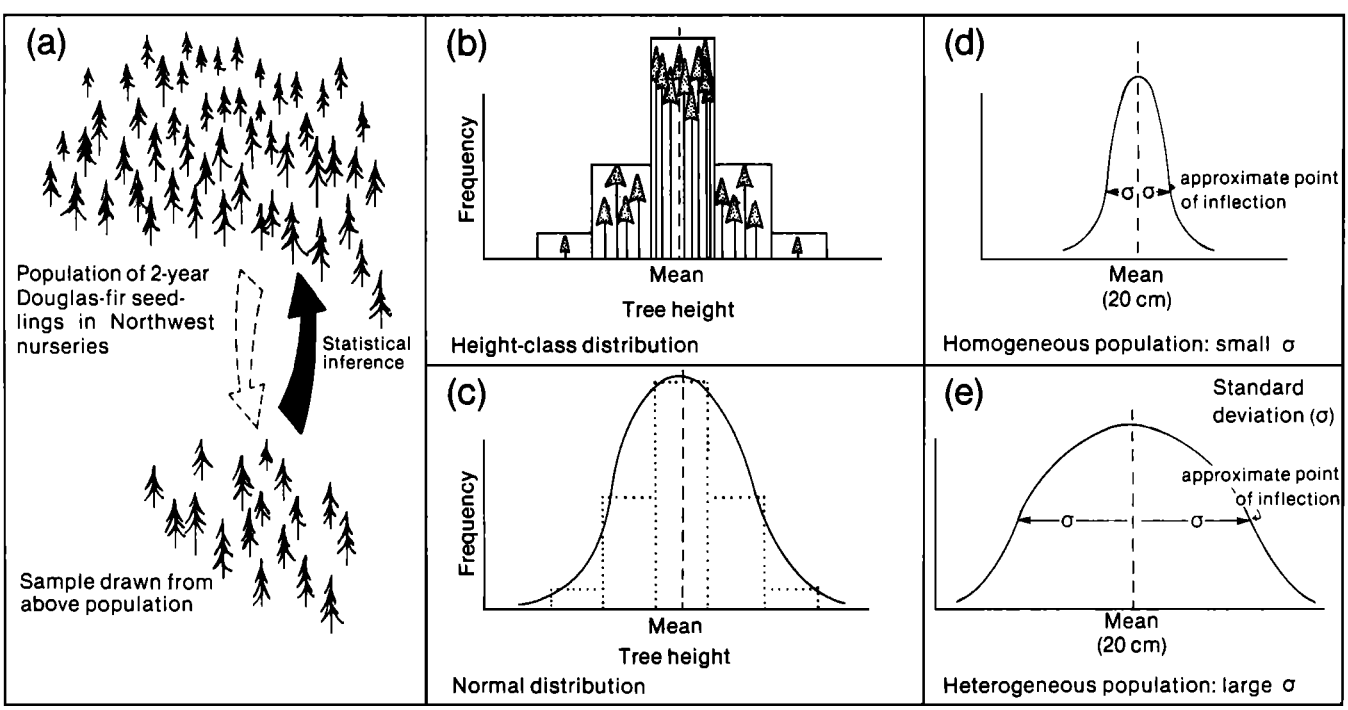


Figure 1. Foresters can use statistical concepts to gain meaningful silvicultural information. For instance, a sample (a) can be drawn from the population of two-year-old Douglas-fir seedlings in Pacific Northwest nurseries. Various sample characteristics can then be determined (b–e) to reveal more about tree growth.

one of several **population parameters.** Because these parameters are virtually impossible to determine exactly because of the vastness of a population, we generally measure a **sample estimate** that corresponds to a particular population parameter. For example, we may select a representative sample of two-year-old Douglas-fir seedlings from Northwest nurseries and try to ascertain mean sample height. Sample estimates are far more useful than population parameters because measuring every individual in a population rarely is feasible. In fact, most statistical conclusions will be based on interpretation of sample estimates, not population parameters. But bear in mind that the final focus reverts to the population *(fig. 1a).*

Knowing only the mean is not enough. We must look at how values are distributed around the mean. The **normal distribution,** or bell-shaped frequency curve *(fig. 1c),* is the symmetrical distribution occurring most often in nature. The curve is represented by many observations near the mean, or center, and few observations near the "tails," or ends. The **standard deviation** characterizes the dispersion (or "spread") of individuals in a population or sample values about the mean (Freese 1967) and is commonly used to "qualify" the mean.

For example, if we estimate the population mean of two-year-old Douglas-fir seedlings to be 20 cm, the more uniform (homogeneous) the population is, the smaller the standard deviation will be. Conversely, the more diverse (heterogeneous) the population is, the larger the standard deviation will be *(fig. 1d).* Geometrically, the distance between the mean and the point of inflection of the normal distribution curve is the standard deviation. Algebraically, the standard deviation is the square root of the **variance.** The more dispersion in a population, the greater the variance. This of course influences sampling—that is, we need a larger sample to adequately test a more diverse population than we do to test a more uniform one.

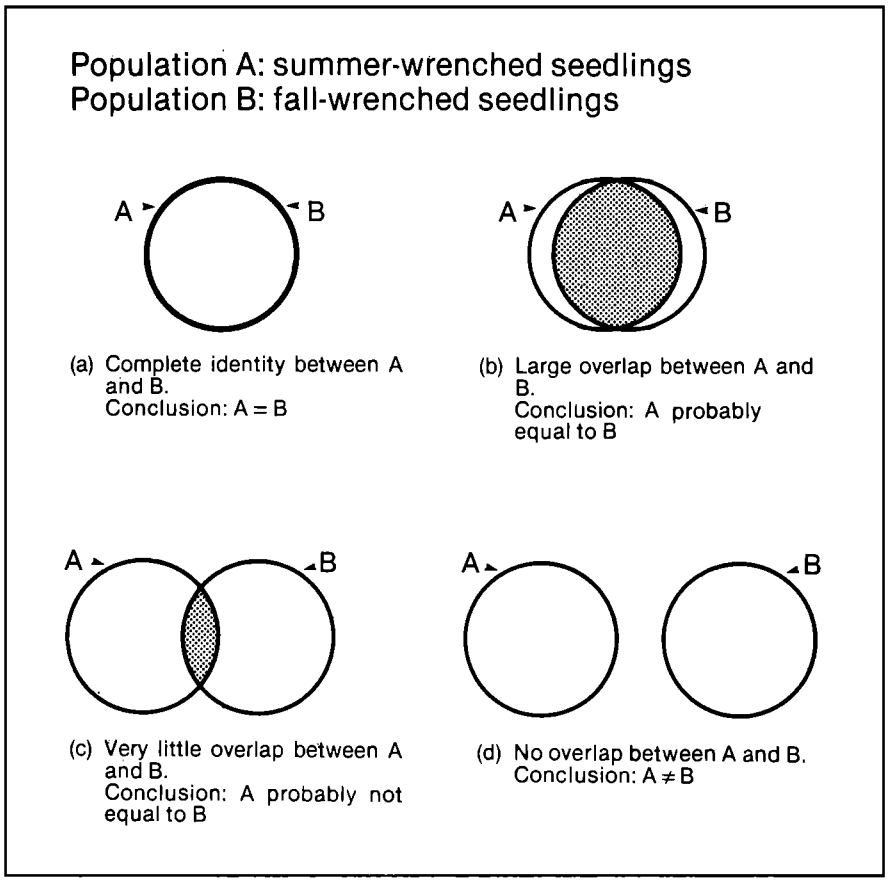


Figure 2. Assessing the similarities and differences of two seedling populations.

The sole purpose in calculating sample estimates as specific values is to get a handle on a biological question. For instance, we may calculate mean sample height of two-year-old Douglas-fir seedlings to better assess the silvicultural implications of using a new nursery practice or site-preparation technique. Whether we ultimately consider that practice or technique operationally feasible will depend on the statistical inferences drawn from our sample.

Hypothesis Testing

Suppose, for example, that a forester wants to compare the results of root-wrenching Douglas-fir seedlings in summer and in fall. The populations of interest would be all summer root-wrenched and all fall root-wrenched Douglas-fir seedlings. The objective is to determine if one population differs from another in some measured characteristic—say, height or diameter growth, survival, or shoot : root ratio.

Hypotheses may be phrased in two ways. The **null hypothesis** states an equality between population parameters—for example, mean height of summer-wrenched seedlings equals that of fall-wrenched seedlings. The **alternate hypothesis** states a difference between parameters (and what the forester hopes to prove)—for example, mean height of summer-wrenched seedlings is less than that of fall-wrenched seedlings. The null and alternate hypotheses are purposely stated this way to provide a basis for proof by contradiction (Mendenhall 1975). If, on the basis of our study, we were able to soundly reject the null hypothesis, then we could conclude that the alternate hypothesis was correct.

The decision would be easy if all experimental results were as evident as those in *figures 2a* and *2d.* In *2a,* populations clearly are equivalent, and in *2d* they are distinctly different. Unfortunately, the situation is usually more like the intermediate cases *(2b* and *2c),* in which there is some uncertainty about population similarities and differences. In *2b* we would probably conclude that summer- and fall-wrenched seedlings are the same because of the large overlap between the two populations. Conversely, in *2c* we would probably conclude that summer- and fall-wrenched seedlings are different because of the minimal overlap. But because statistics

"Thou shalt not sit with statisticians nor commit a social science."

can never give a definite answer and can only quantify the shade of grey, we will occasionally reach the wrong conclusion—that is, make errors.

Statistical errors are of two types *(fig. 3)*. For instance, suppose the null hypothesis (H_0) states that summer-wrenched seedlings are equivalent to fall-wrenched seedlings. If we concluded in *figure 2c* that treatment effects of summer and fall wrenching were different when in fact they were the same, we would be making a **Type I error** *(fig. 3)*. The **probability** of making a Type I error is denoted by α.

On the other hand, if we concluded in *figure 2b* that treatment effects of summer and fall wrenching were the same when they actually were different, that would be a **Type II error** *(fig. 3)*. The probability of making a Type II error is denoted by β. Unfortunately, Type I and II errors are not independent of one another and often work against each other. The best that experimenters can hope to do is minimize the probability of making these errors.

Statistics can help in the common situations like *figures 2b* and *2c* by allowing us to quantify (in terms of probability) the risk taken when we conclude that summer- and fall-wrenched seedlings are the same in *2b* but different in *2c*. The term **significant** is often used in this context. "Significant" really means "significantly different"—that is, significant results show a real difference between populations (for example, between summer- and fall-wrenched seedlings). The **significance level** refers to the probability that we have drawn the wrong conclusion and is defined as α—that is, the probability of making a Type I error. Traditionally, an acceptable α level is 0.05 or 0.01. If, for example, in *figure 2c* we conclude that summer- and fall-wrenched seedlings differ significantly at $\alpha = 0.05$, there is a 5-percent chance that we are wrong (and a 95-percent chance that we are correct).

The α, or significance, level reflects the amount of confidence the experimenter has in a conclusion; a small α corresponds to a high degree of confidence because the chance of being incorrect is relatively small. To choose the appropriate α level, the experimenter should estimate the consequence that would be incurred by concluding that summer- and fall-wrenched seedlings are different when they are not. If such a mistake would be costly, the α level should be set quite small so that a Type I error may be minimized. This is something only the experimenter can decide. Because statistics allows us to draw a conclusion without being absolutely certain, we must always assign that conclusion a probability of being incorrect (in this example, 5 percent); otherwise it might be misleading.

It is important to stress, however, that a test that is statistically significant cannot tell us if the observed difference is important silviculturally or biologically. It can only tell us that the observed difference is probably not caused by chance, or natural variation, alone.

Comparing Means

The root-wrenching case we have been discussing involves the comparison of two sample means. Suppose the mean height of a sample of summer root-wrenched seedlings was 30 cm and that of fall-wrenched seedlings was 45 cm. We need to know if that 15cm height difference is "large" or "small" to determine whether this is a true treatment difference or one due to chance, or natural variation, alone. If we assume that each sample was relatively small, that it was **randomly** selected, and that the measurements came from normally distributed populations (see *fig. 1c*), then the t-test is the appropriate statistical tool. The **t-test** is conducted by dividing the difference between the two sample means by an appropriate estimate of the variation (dispersion).

Variation among sample means is due partly to **sampling error** (the variation, due to chance alone, incurred by selecting random samples to represent the population). Sampling error will always be present because each sample mean cannot exactly equal the population mean. Nonsampling errors (inaccuracies from, say, poor field technique or calibration mistakes) can also account for variation among sample means.

In the root-wrenching case, by dividing the difference between the two

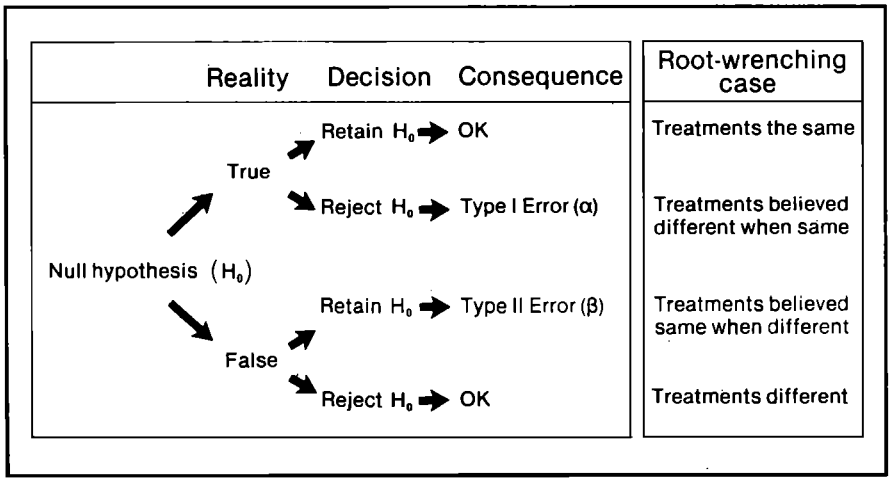


Figure 3. Decision-making and its possible consequences. The null hypothesis (H_0) states that there is no difference between summer- and fall-wrenched seedlings. A Type I error would result if we concluded that seedlings from the two wrenching treatments were different when they were the same. A Type II error would result if we concluded that seedlings were the same when they were different.

means (15 cm) by an estimate of the variation, we can scale that difference and use our estimate of the variation as a yardstick of the amount of variation due to chance alone. This helps us decide statistically if the 15cm difference is large or small. If 15 is large relative to our estimate, then there is probably a real difference between the summer- and fall-wrenched seedlings. If 15 is small relative to our estimate, then the difference is probably due to chance alone (i.e., sampling error). In a population where data are dispersed (see *fig. 1e*), 15 units may be a relatively small difference, but in a population where data tend to cluster (see *fig. 1d*), 15 units may be a relatively large difference. This emphasis on relative, rather than absolute, values is the hallmark of statistics over mathematics.

Analysis of Variance

Suppose a forester wants to compare four root-wrenching treatments (bimonthly intervals from April through October) and a **control.** Note that a control can be considered either separately or as another treatment. In this case the forester would be testing a hypothesis about five populations, not just two. This case could be handled with the popular and commonly used technique **analysis of variance (ANOVA).**

It would be far less misleading if this technique were called analysis of means (Iverson and Norpath 1976) because it tests for differences between two or more treatment means. In fact, an analysis of variance testing the difference between only two means is actually the same as the t-test just discussed. Using ANOVA, foresters can compare the variation between different root-wrenching treatment means with the amount of variation inherent within the experimental seedlings themselves.

To get a clearer picture of this, let us present our t-test root-wrenching example as an ANOVA problem, again assuming randomly selected samples drawn from normally distributed populations. We are comparing the variation between the two treatment groups (summer- and fall-wrenched seedlings) with the variation inherent within each treatment group. If mean sample height differs more between treatment groups than within either one, then the two populations probably truly differ. That is, the variation is too large to have been caused by sampling error (chance) alone. But if mean sample height of the two treatment groups is about the same, then the populations probably do not truly differ. That is, the difference between the means is small enough to have occurred by chance. Even if the mean height of the fall-wrenched seedlings is just slightly larger than that of the summer-wrenched seedlings, we would conclude that sampling error alone could account for this difference and, therefore, that the two treatment groups are not significantly different.

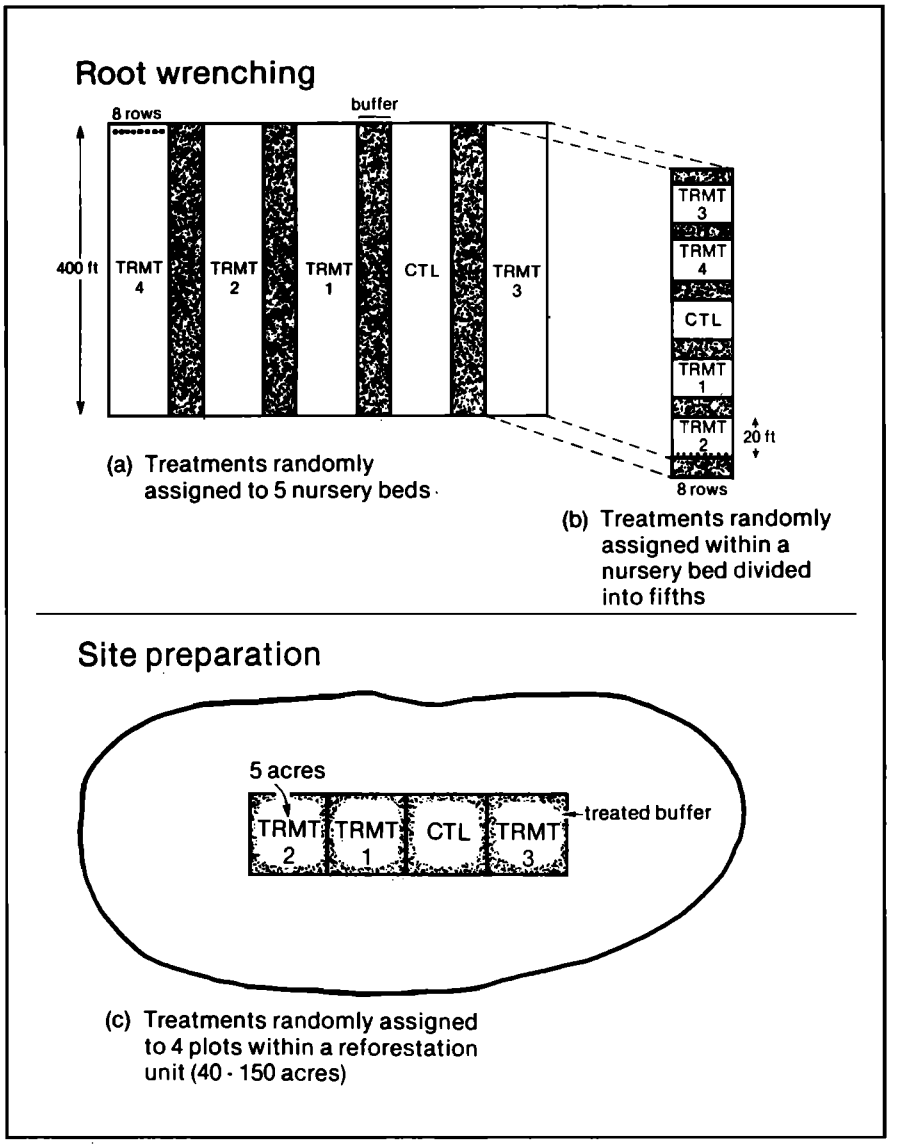


Figure 4. Random assignment of treatments (TRMT)—including a control (CTL)—(a) to five nursery beds, (b) within one nursery bed divided into fifths, and (c) to four plots within a reforestation unit.

Experimental Design

Randomization and replication are the basic principles of sound **experimental design** (Little and Hills 1978).

Randomization can be likened to an insurance policy (Cox 1958): it helps guarantee that no one treatment is preferentially assigned to an **experimental unit.** Randomization ensures a valid measure of **experimental error.** We would not want to jeopardize the credibility of the total experiment by, for example, assigning a favorite treatment to the healthiest seedlings—

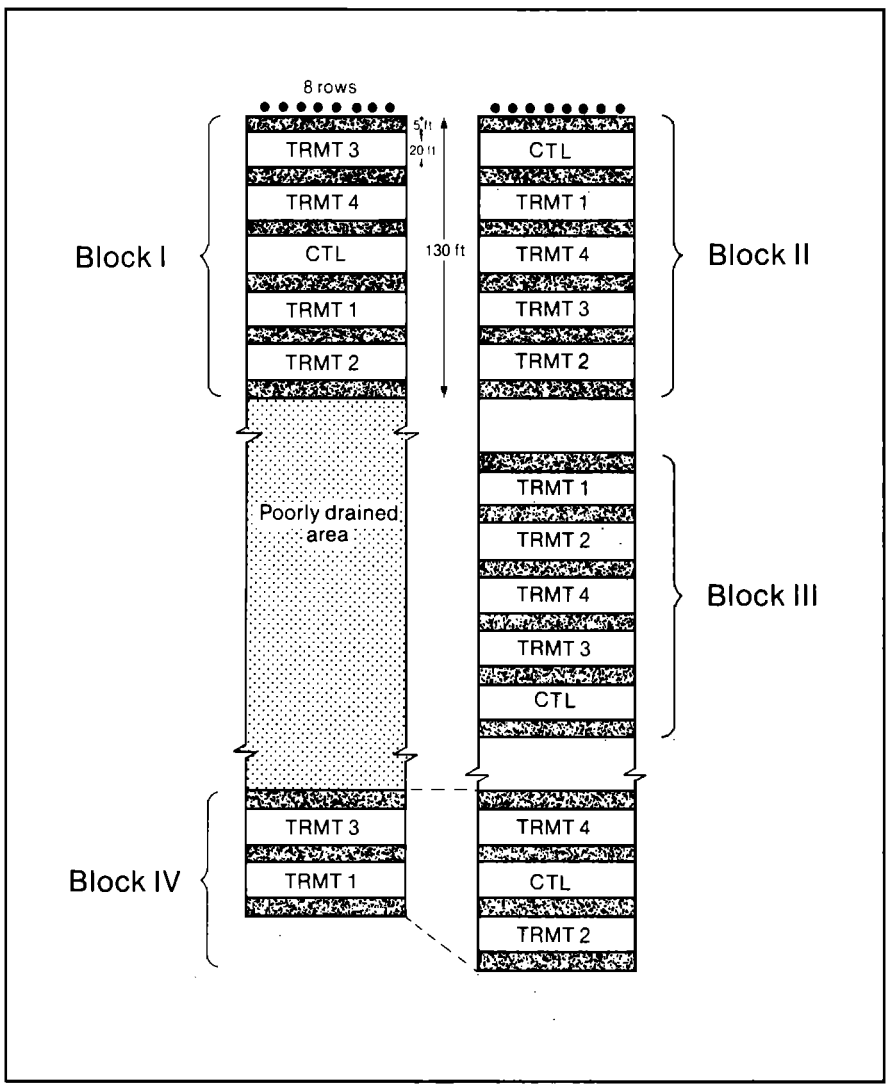


Figure 5. Root-wrenching study in which five treatments (TRMT = treatment, CTL = control) are randomly assigned within each of four blocks established in two nursery beds.

which probably would have grown the most anyway—or outplanting trees on a prepared site that had less slash and more mineral soil, which might give those trees an innate advantage regardless of the treatment applied.

In the root-wrenching case, we can randomly apply the five treatments (including a control) to all seedlings in each of five nursery beds *(fig. 4a)*, or we can divide one nursery bed into five equal lengths and randomly assign the five treatments to each fifth, if using five beds is not operationally feasible *(fig. 4b)*. Similarly, in the case of site preparation, we could divide each of four reforestation units into four 5-acre plots and randomly assign one of the four site-preparation treatments (including a control) to each plot *(fig. 4c)*. Notice in *figure 4c* that the site-preparation treatments are contiguous. They do not have to be: this is only one possible configuration of the experiment. The topography and local conditions usually will determine the physical arrangement.

Replication is the repetition of a treatment on more than one group of seedlings. This provides a way of computing experimental error. Without replication, we have only a case study—an unreplicated experiment with limited applicability, or scope of inference. There is no way of knowing if results are significant or if they can be reproduced.

What we are replicating is often confusing. However, there is an important distinction between measuring 2,000 seedlings in one treated nursery bed and measuring 500 seedlings in each of four similarly but independently treated beds. Even though the same total number of seedlings would be measured in both cases, the latter case provides stronger results because we have truly replicated the treatments, not just the **sampling units** (seedlings).

A group of experimental units to which a complete set of treatments is assigned is called a **block.** For example, the group of five nursery beds in *figure 4a* and the single bed divided into fifths in *4b* both are blocks because each contains a complete set of five treatments. Similarly, each reforestation unit containing all four treatments *(fig. 4c)* is a block. Blocks can be areas of different soil type, slope, aspect, vegetative competition, or any other characteristic or combination of characteristics.

To illustrate blocking, suppose we designed an experiment to determine how the previously mentioned bimonthly wrenchings would affect seedling morphology. We will use a uniform portion of a nursery bed as a block and would like four blocks in total *(fig. 5)*. Each block, 130 feet long, is to comprise five 20-foot treatment areas and 5-foot buffer zones between these areas. **Buffers** are usually untreated areas between treatments where plots are small (see *fig. 4a, b*) but can actually be part of the treated area where plots are large (see *fig. 4c*). But such treated

"The notion that we can prove anything by manipulating numbers with statistics is a popular misconception."

buffers are not actually sampled. Because only two beds are available, each must accommodate more than one block.

After walking the beds, we find that they are not uniform. Though we can easily find room for two blocks in one bed, the other has several hundred feet of poorly drained soil that makes it undesirable for use in this study. Nonetheless, the last 60 feet of both beds are in a higher lying area and are healthy. Differences within blocks should be as small as possible so that treatment location within the block will not bias treatment response. However, differences between blocks should be as large as possible so that the same treatment, applied in each block, has an equal chance of performing under a variety of experimental conditions. Therefore, blocks should not be assigned randomly within an experiment. Since the blocking criteria are to make the differences between blocks as large as possible but the differences within blocks as small as possible, we can consider the last 60 feet of both beds, combined, as block IV.

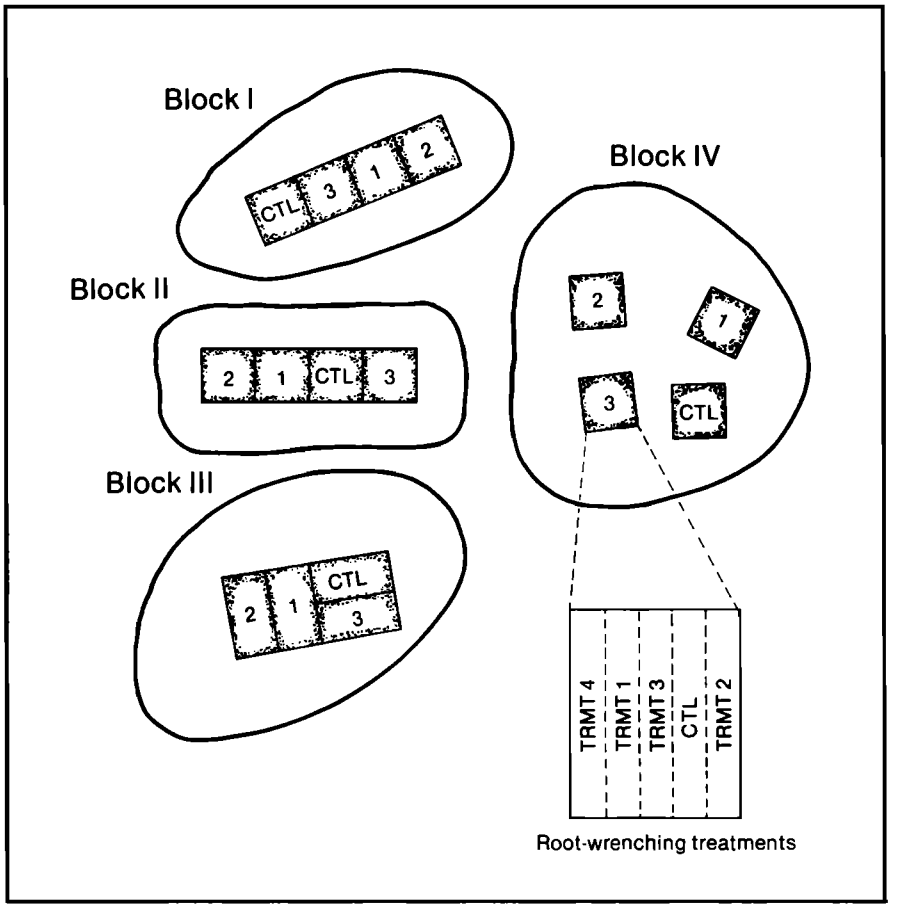


Figure 6. Split-plot design for the outplanting phase of a nursery study. Nursery root-wrenching subplot treatments (TRMT) and a control (CTL) are superimposed on the site-preparation whole plot treatments (1 = spray, 2 = slash and burn, 3 = scarification with bulldozer) and a control on four reforestation units.

Design Types

The basic principles that we have discussed apply to all types of experimental designs. The design merely tells the experimenter how to arrange the treatments within the replications.

For example, in a completely randomized design, no consideration is given to the physical arrangement and proximity of individual treatments. In a randomized complete block design, a complete set of all treatments is grouped together in a block, with treatments randomly assigned within the block, as in our root-wrenching and site-preparation cases *(figs. 4 and 5)*. In a split-plot design, two sizes of experimental units—the whole plot and the subplot—are combined, the subplots superimposed on the whole plots. The treatments applied to the whole plots need to be applied to a larger area than the treatments applied to the subplots. Many other types of designs exist. Those named here are probably the most often used by foresters.

Planning

Many problems can be avoided if foresters carefully plan the stages of experimentation and formulate them into a study plan written before beginning an experiment. For example, conducting an outplanting phase of a nursery study is often desirable so that field performance of seedlings grown under various nursery cultural practices can be evaluated. By combining our root-wrenching and site-preparation cases, we can examine each study individually and also look at any interactions between the various wrenching and site-preparation treatments. Let us consider how this combination could be arranged in a split-plot design.

Suppose we want to install a set of four site-preparation treatments (CTL = control, 1 = spray, 2 = slash and burn, 3 = scarification with bulldozer) on each of four reforestation units (blocks) *(fig. 6)*. Again, notice that the 5-acre plots within each block do not need to be contiguous as long as the area over which the treatments are applied within the block is as uniform as possible.

We planned to have 200 seedlings per block available for the outplanting

phase of the nursery experiment. We could divide those 200 seedlings into four groups of 50 and arrange each group of 50 in five rows of 10 seedlings each (each row representing a different root-wrenching treatment, randomly assigned) on each 5-acre plot of each site-preparation block *(fig. 6)*. That is, we could superimpose the nursery outplanting study on the existing site-preparation study, thereby creating a split-plot design in which the larger site-preparation plots are the whole plots and the small rows of root-wrenched seedlings, 50 per plot, are the subplots. This would allow us to test for differences among the site-preparation treatments and the nursery root-wrenching treatments, and for interactions between the two sets of treatments.

In the split-plot illustration just described (see *fig. 6*), it is critical to decide how many seedlings to plant initially so that adequate numbers are produced for all phases of the total study. Experimenters must take into account the expected survival (or mortality), the desired precision of measurement, the study duration, the kind of sampling, and the number of measurements to be taken.What to measure is often a problem, but it must be determined before designing an experiment to ensure adequate sample size. Survival cannot be meaningfully measured on just one seedling. We would need 10 seedlings to measure survival to the nearest 10 percent and 100 seedlings to measure it to the nearest 1 percent. But if resources limit foresters to measuring survival to the nearest 5, 10, or 20 percent, then they must consider ahead of time the consequences of a reduced sample size. Sometimes the best decision is not to do the experiment at that time and to wait until more resources are available.

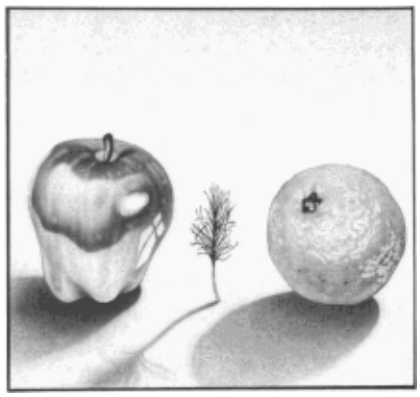
Working Statistics

The proof of how well a forester uses statistics is how well the experiments answer the questions they are designed to address.

The split-plot illustration described above points up the strength and flexibility of installing soundly planned experimental designs. Requesting help from a knowledgeable source—a statistician or biometrician—*before* data are collected can be a valuable step in preventing waste and frustration during a study. Similarly, the type of analysis to be conducted should be determined in the early planning stages. Since the advent of computers, the wrong answers can now be reached much more quickly and easily than ever before. Remember that the computer, like statistics, is only a tool and must not be allowed to dictate experimental or analytical approach.

Moreover, many excellent commercial software packages (programs) are available—SAS (Barr et al. 1976), SPSS (Nie et al. 1970), SPSS-X (SPSS Inc. 1983), BMDP (Dixon 1983), and MINITAB (Elkins 1971), to mention just a few. If one package is inadequate, investigate others. A wide variety of techniques, in addition to the t-test and ANOVA, is available for analyzing data. Had we been interested in the relationship, say, between seedling growth and plant moisture stress, regression analysis could have been applied. Foresters should consult general statistics texts or a statistician or biometrician for the required details.

In sum: Always keep in mind the objectives and purpose in conducting an experiment; never lose sight of the big picture for the sake of small details. Incorporate replication and randomization where appropriate and do not hesitate to request help from qualified sources. Refer to general statistics texts such as Little and Hills (1978), Freese (1967), Steel and Torrie (1981), and Mendenhall (1968, 1975); these texts, which cover the techniques discussed in this article in greater detail, are very readable. With the basic tenets of experimental design in mind, foresters should be well on the way to planning statistically sound experiments that produce statistically valid results. ■



Literature Cited

BARR, A.J., J.H. GOODNIGHT, J.P. SALL, and J.T. HELWIG. 1976. A User's Guide to SAS '76. 320 p. SAS Institute, Raleigh, NC.

CLARKE, G.M., and D. COOKE. 1978. A Basic Course in Statistics. 368 p. A Halsted Press Book, John Wiley & Sons, NY.

COX, D.R. 1958. Planning of Experiments. 308 p. John Wiley & Sons, NY.

DIXON, W.J. 1983. BMDP Statistical Software. 733 p. Univ. California Press, Berkeley.

ELKINS, H. 1971. MINITAB Edit, MINITAB Frequencies and MINITAB Tables: A Set of Three Interrelated Statistical Programs for Small Computers. 100p. Univ. Chicago, IL.

FREESE, F. 1967. Elementary Statistical Methods for Foresters. 87 p. U.S. For. Serv. Agric. Handb. 317.

HUFF, D. 1954. How to Lie with Statistics. 142 p. W.W. Norton & Co., NY.

IVERSON, G.R., and H. NORPATH. 1976. Analysis of Variance. 95 p. Sage Publications, Beverly Hills, CA.

KIMBLE, G.A. 1978. How to Use (and Misuse) Statistics. 290 p. Prentice-Hall, Inc., Englewood Cliffs, NJ.

LITTLE, T.M., and F.J. HILLS. 1978. Agricultural Experimentation: Design and Analysis. 350 p. John Wiley & Sons, NY.

MENDENHALL, W. 1968. Introduction to Linear Models and the Design and Analysis of Experiments. 465 p. Duxbury Press, Wadsworth Publishing Co., Inc., Belmont, CA.

MENDENHALL, W. 1975. Introduction to Probability and Statistics. 460 p. Duxbury Press, North Scituate, MA.

NIE, N.H., C.H. HULL, J.G. JENKINS, K. STEINBRENNER, and D.H. BENT. 1970. Statistical Package for the Social Sciences. Ed. 2. 675 p. McGraw-Hill Book Co., NY.

SPSS INC. 1983. SPSS-X: User's Guide. 806 p. McGraw-Hill Book Co., NY.

STEEL, R.G.D., and J.H. TORRIE. 1981. Principles and Procedures of Statistics: A Biometrical Approach. 633 p. McGraw-Hill Book Co., NY.