Chapter 28
Designing Nursery Experiments

T. L. White

Abstract

Nursery managers face a wide variety of problems that lend themselves to research methods. This chapter (1) describes fundamental statistical concepts—inference, replication, randomization; (2) discusses methods for controlling experimental error—pairing, blocking, split-plot design, covariance—to increase the sensitivity of experiments; and (3) traces the research process applied to forest-tree nurseries-defining problems, designing and conducting experiments, and implementing solutions based on sound Interpretation of data. Combining the statistical concepts presented here with personal experience and biological intuition strengthens the nursery researcher's ability to meet the major goal—and challenge—of nursery research: to develop new methods for producing high-quality seedlings at low cost.

28.1 Introduction

Nursery managers and growers are researchers both by need and inherent nature. Keenly observant and inquisitive, they continually seek to improve seedling quality and cost-effectiveness of their nursery practices. They face a wide variety of problems that lend themselves to research methods—for example, whether to use a new piece of equipment or a new herbicide, how dense and when to sow seed for various stock types, or how to determine optimum fertilizer and irrigation regimes. Necessarily, however, recommendations to alter nursery practices are nearly always based on incomplete information, which successful nursery managers evaluate in light of their experiences and instincts to make sound, effective decisions. The science of **statistics** deals with drawing conclusions from incomplete information, whereas **biometrics** is the application of these statistical techniques to biological problems. Statistically designed experiments often provide important information upon which nursery managers can base their decisions and calculate the degree of uncertainty associated with their conclusions.

The design and execution of experiments are often team efforts involving biometricians and researchers. Information in this chapter should help make the nursery researcher a stronger team member, better able to balance statistical considerations with practical and biological aspects of nursery problems. Specifically, the chapter objectives are to (1) trace the research process as applied to forest-nursery problems (see 28.3), (2) contrast operational trials with 'statistically designed experiments (see 28.4), (3) describe, intuitively, the statistical concepts of designed experiments (see 28.5), and (4) delineate the processes involved in designing, executing, analyzing, and interpreting those experiments (see 28.6).

In Duryea. Mary L., and Thomas D. Landis (eds.). 1984. Forest Nursery Manual: Production of Bareroot Seedlings. Martinus Nijhoff/Dr W. Junk Publishers. The Hague/Boston/Lancaster, for Forest Research Laboratory, Oregon State University. Corvallis. 386 p.

28.2 How to Use This Chapter

A few words are in order regarding use of this chapter.

The chapter is intended as a complete introductory treatise on experimental design for nursery workers; no statistical training is assumed. It is not a formula-oriented discussion; excellent texts by Freese [8] and Little and Hills [13] provide formulas for analyzing basic experimental designs. Nor does it specifically address the plantation-testing phase of nursery research, although the statistical concepts described may be widely applied to agricultural and forestry experimentation.

The chapter may be useful both for a first-time reading, to gain an understanding of statistical concepts of experiments, and for future reference. Some concepts are treated more fully than needed for a first reading and should probably just be skimmed. In particular, a quick reading of definitions (Table 1) and sections 28.3 to 28.5 is good preparation for section 28.6, the main focal point, describing how to design and execute experiments and interpret their results. For future reference, section 28.6 can be used as a checklist of items to consider when actually planning nursery experiments.

28.3 The Research Process: An Overview

An **experiment** is a planned inquiry to obtain new information or to confirm or deny previous results for the purposes of making recommendations [18]. The process of experimentation is profitably applied to many problems encountered in forest-tree nurseries (Fig. 1). Cause-effect relationships [19, p. 86] are established by observing how certain **response variables** (say, caliper) are influenced by specified levels of one or more **factors** (say, fertilizer). Although nursery managers may not necessarily think of it as experimentation, they commonly (1) encounter a problem, (2) seek out existing information, (3) refine the problem and set hypotheses or objectives, (4) plan and conduct experiments to obtain new data pertinent to their nursery conditions. (5) draw conclusions based on their interpretation of the new data in light of existing information and their instincts, and (6) implement a change in nursery procedure. Often, because conclusions lead to new problems or questions, several stages of experimentation may be required. (See also chapter 29, this volume, for more details on problem-solving techniques.)

Table 1. Definitions of common terms used in the design of experiments.

Term	Definition	Examples/Comments
Experiment	Planned inquiry to obtain new information or to confirm or deny results from previous investigations for the purposes of making recommendations [18, p. 88].	Nursery experiments gain information on new field techniques, equipment, packing-shed alignments, storage facilities, etc.
Operational trial	Preliminary experiment in which each treatment is applied to only one plot (nonreplicated).	Useful when treatment effects will be large relative to background "noise" (uncontrolled variation).
Designed experiment	Detailed, critical investigation in which precise, unbiased conclusions and measures of uncertainty associated with those conclusions are required.	Treatments are *nearly always* replicated more than once on separate plots and allocated to plots at random.
Inductive reasoning	Drawing conclusions or making predictions about a wide sphere of interest (a population) from particular cases or observations (samples).	The sun has risen every day for millennia (a large sample); therefore, it will rise tomorrow.
Deductive reasoning	Drawing conclusions or making predictions based on well-defined principles from which those conclusions or predictions logically follow.	That the sun will rise tomorrow logically follows from the principles of astronomy.
Factor	An item, element, or process under investigation in an experiment. Effects of a given factor are examined by testing each factor at more than one level (factor level).	Sowing date, irrigation, seed source, bed density, etc. are factors; three rates of nitrogen (N1, N2. N3) and two sowing dates (D1, D2) are factor levels.
Treatment	All factors and their levels applied to an experimental plot.	From above, N1/D2 is a treatment plot sown on the second date receiving the lowest nitrogen level.
Experimental plot	Smallest physical unit to which a treatment is allocated independent of all other treatments.	A specific length of nursery bed.
Observational unit	Observed or measured items within an experimental plot [11, p. 9].	Tree seedlings within a nursery experimental plot.
Measurement plot	Portion of the experimental plot actually measured; unmeasured portion serves as buffer or border.	The center of a nursery plot; seedlings on either side of center are *not* measured.
Response variable	Variable (characteristic) measured on each experimental plot to assess influence of treatments.	Number of plantable seedlings, percent germination. height, caliper, shoot:root ratio.
Precision	Relative dispersion or clustering of measurements or estimates.	A precise measurement is one of low dispersion; if remeasured, it will be nearly the same.
Accuracy	Absolute correctness of measurements or estimates.	A scale that *always* weighs 2 g too heavy is inaccurate even if precise (consistent).
Bias	Directional (up or down) measure of inaccuracy.	The above scale is biased upward 2 g.
Confounding	Condition in which the effects of two of more factors on a given response variable are confused and cannot be separated.	A nursery researcher finds larger seedlings from the field with both higher N and P levels; the effects of N and P *cannot* be separated.
Experimental error	A measure of the variation among experimental plots receiving identical treatments [18, p. 901.	Experimental error will be high if field plots are inherently variable or if experimental technique is sloppy.

For example, perhaps during an initial experiment, a new rotating root-pruning table is found to save money on the packing line, but seems to damage too many roots. How can the table be redesigned to hold the seedlings in place better so fewer roots are damaged? Existing information may be sought from engineers and manufacturers, and objectives formulated for designing one or two modifications to test in a second experiment. Having then tried the new modifications for a period of time, the manager decides that one of them "causes" the desired "effects" (less root damage) and operationally implements the new root-pruning procedure on the packing line.

Consider, as a further example, the process of experimenting with a new herbicide to control weed species in forest nurseries [16]. Experiments are set up to compare various application rates of a new chemical to the standard weed-control method to determine relative phytotoxicity and effectiveness. Note that *no* amount of experimentation can totally prove beyond all doubt that this chemical will be suitable for all nurseries under all conditions. This is common of problems requiring **inductive reasoning—**in which inferences are made about a larger sphere of interest from a smaller data base. The broader and more intensive our **sample—**that is, the more years and nurseries in which the chemical is used and tested—the more comfortable (certain) we feel about applying the results to the **population of interest,** perhaps all nurseries of similar soil type. We further use **deductive reasoning,** based on underlying biological, chemical, and physical principles, to extend and rationalize the inductive inferences drawn from experimentation. For example, the new chemical will probably not be particularly suitable for nurseries suffering severe grass competition if it is chemically ineffective against grasses.

Because experimental processes do not absolutely prove the hypotheses being investigated, the amount of data (sample size) required to make a decision becomes a personal choice. How certain must the conclusions be? Some problems require stronger evidence than others, perhaps because the decision would be more costly to implement. Thus, the type and amount of experimentation and the formality with which the research is applied to a problem vary with both the investigator and the problem.

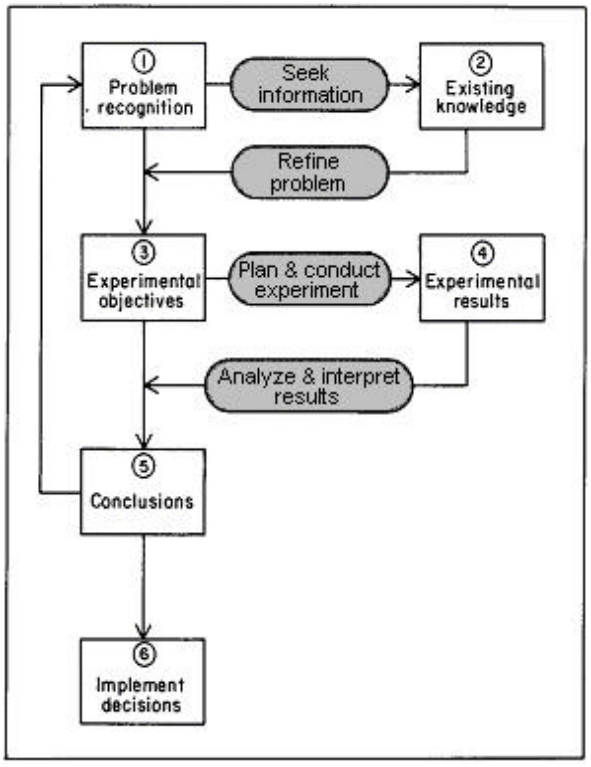

Figure 1. The research process, as it might be applied in forest-tree nurseries (adapted from [19, p. ix]).

28.4 Operational Trials versus Designed Experiments

Some problems are appropriately addressed with **operational trials** (preliminary investigations in which each treatment is applied to only one plot). Others must be examined through **designed experiments** (detailed, critical investigations in which precise, unbiased conclusions and associated measures of uncertainty are required). The nature of yet other problems precludes *any* type of experimentation. For example, when a manager discovers a widespread insect or disease problem in the nursery, immediate control by the "recommended" method is more important than experimenting to find the optimum rate or chemical. The manager may decide to experimentally treat some small areas at a lower rate or with an untried chemical, but even this may be unacceptable because of the large inoculum source remaining if the untested **treatment** were unsuccessful.

28.4.1 Operational trials

In general, operational trials have two main uses: preliminary investigations and final-phase, large-scale testing. Operational trials are particularly suitable for preliminary experiments when background variation among **experimental plots** is small relative to expected treatment effects. For example, the effects of a new herbicide formulation on weed mortality will be large relative to the background effects due to other causes. If the preliminary objectives are to see whether the chemical has any promise and warrants further testing, an operational trial is suitable for visually assessing effectiveness and phytotoxicity. Many preliminary (screening) tests of, for example, new chemicals, fertilizer regimes, wrenching blades, and packing-line arrangements are appropriately conducted as unreplicated operational trials.

For final-phase or large-scale testing, the use of small experimental plots needed to obtain sufficient replication may generate experimental artifacts. That is, the new treatment *cannot* be "operationally" applied to the small experimental plots often needed for designed, well-replicated experiments. Operational trials are suitable in these situations. Larger plots, more representative of operational application of the new treatment, are used, while statistical replication is sacrificed.

When an operational trial is chosen as the appropriate method of experimentation, the various treatments should be (1) applied to similar areas to minimize systematic effects due to uncontrolled variation and (2) compared in more than one nursery bed so that the inferences and conclusions drawn will have broader validity.

28.4.2 Designed experiments

Given that an experiment is warranted, a designed experiment is more appropriate than an operational trial for cases benefiting from one or more of the special attributes of designed experiments (Table 2). Designed experiments are particularly useful for detailed investigations—for example, establishing optimum rates of a chemical or procedure, investigating interactions among multiple factors, or revealing the biological principles of a phenomenon under investigation. In these cases, the attributes of designed experiments are well worth the extra effort. Uncontrolled, background variation affecting growth rate within and between nursery beds is large due to differences in fertility, soil type, water drainage, and irrigation.

Well-designed experiments, randomized and replicated over this background variation, are often required to achieve unbiased estimates of treatment response with the appropriate level of precision and range of validity. Most of the remainder of this chapter considers the concepts behind and execution of designed experiments in nursery research.

28.5 Statistical Concepts of Designed Experiments

Statistical-inference concepts, discussed first (see 28.5.1), center around developing and quantifying the uncertainty of the conclusions drawn from designed experiments. Replication and randomization, discussed next (see 28.5.2), are the two core concepts of experimental design and are considered in some depth; many of the benefits of designed experiments derive because they are, by definition, randomized and replicated.[1] Finally, methods for controlling experimental error (see 28.5.3) are presented which can result in more precise estimates of treatment means and more certainty about the conclusions drawn from experiments.

28.5.1 Statistical inference

28.5.1.1 Point estimates, interval estimates, and hypothesis tests

Statistical inference is the process of using sample data to generalize about a population or wider sphere of interest. For designed experiments, this usually implies calculating the level of uncertainty associated with these generalizations. This is one of the main rationales for using designed experiments.

An example should clarify the three main elements of statistical inference (Table 3) and the importance of these in nursery experimentation. If five plots receive a specific fertilizer regime, the mean number of plantable seedlings for that regime would be calculated by summing the total number of plantable seedlings for all five plots and dividing the sum by 5. This treatment mean is subject to experimental error and is only a **point estimate** of the true population mean (the response achieved if that regime were applied to an infinite number of nursery plots). Though further experiments might show this point estimate to be in error, for now, it is a single number that estimates the parameter of interest.

[1] Nonreplicated and fractionally replicated designs are of limited usefulness in nursery research and are not considered here; see Cox [7] and Kempthorne [11].

Confidence intervals quantify the uncertainty and state the error associated with point estimates. Note that the span of a confidence interval is closely related to the experimental error. If the just-mentioned fertilization experiment yields highly reproducible results (i.e., precise, low experimental error), then the experimenter can state with a high degree of confidence (say, 95%) that the true mean fertilizer response lies within a narrow range surrounding the estimated treatment mean.

Hypothesis testing allows researchers to quantify the uncertainty with which they accept or reject hypotheses formulated before an experiment. A statistical hypothesis is testable by experimentation in the sense that experimental results will either tend to support or refute it; however, because it can never be totally proven or disproven, researchers calculate the level of confidence placed on the decision to acceptor reject.

Statistical hypotheses are formulated as **null hypotheses**—that is, the effects under investigation are assumed to have *no* effect on the response variable. Examples are (1) all fertilizer regimes yield the same number of plantable seedlings, (2) bed density has no effect on stem caliper, (3) the effect of nitrogen (N) fertilization on height growth is the same regardless of the level of phosphorus (P) fertilizer (no interaction). This "innocent-until-proven-guilty" approach has both statistical and scientific underpinnings. From a statistical standpoint, for example, a researcher calculates the probability of the observed differences between two treatment means occurring by chance if, in fact, there are no differences between the true treatment effects. If there is only a small chance (say, 5%) of obtaining the observed differences if the null hypothesis is true, the researcher concludes that treatments differ—and rejects the null hypothesis. From a scientific standpoint, null hypotheses state a skepticism and wariness of the consequences of being wrong. For instance, a nursery researcher may not want to implement a new, more costly fertilizer regime until the evidence points overwhelmingly in its favor; that skepticism is maintained by hypothesizing no effect.

28.5.1.2 Incorrect conclusions from experiments

Because statistical hypotheses can never be proven or disproven, it is inevitable that incorrect conclusions are drawn from experiments. Two types of incorrect decisions (errors) are possible (Fig. 2):

- **Type 1 error** (a): The null hypothesis is rejected when it is actually true. That is, differences among treatments are declared statistically significant when the true treatment effects are, in fact, identical.

Table 2. Attributes of designed experiments (adapted from [7, p. 5]).

Attribute	Explanation	Comments/Examples
Absence of systematic effects	Treatment comparisons are not confounded or biased due to uncontrolled (background) variation.	Comparison of two fertilizer regimes would be biased by consistently applying one regime to plots in a more fertile part of the nursery.
Proper degree of precision	Poor design and large, uncontrolled variation result in large experimental error and imprecise estimates of treatment effects; "overdesign" results in overexpenditure of effort for the necessary data.	High precision occurs when (1) experimental plots have similar background characteristics, (2) experimental procedures are conducted with care and accuracy. (3) a large number of replications are used, and (4) the experimental design is efficient [7, p. 154].
Wide range of validity	Inferences and conclusions will apply to the entire population of interest; replication over time and space broadens the range of validity.	Testing a herbicide for a few years in several nurseries results in broadly applicable conclusions.
Quantification of degree of uncertainty	The "reasonable shadow of a doubt" accompanying experimental conclusions is quantifiable.	There is a 99% chance that the new fertilizer regime results in 2+0 height increase of 3.1 to 5.2 cm above the standard regime.

Table 3. Three main elements of statistical inference.

Term	Explanation	Comments/Examples
Point estimate	A single number that estimates a certain quantity in the population of interest.	Treatment mean, standard deviation, minimum, maximum, and range are all point estimates.
Confidence interval	For a given level of confidence, the specified range within which the quantity of interest lies.	A confidence interval on a treatment mean states with, say, 95% confidence that the true mean response lies between two estimated values.
Hypothesis testing	A statistical technique to accept or reject a hypothesis formulated before the experiment in light of the empirical results.	Designed experiments allow quantification of the level of uncertainty associated with acceptance or rejection.

- **Type 2 error (b):** The null hypothesis is accepted when it is really false. That is, statistically significant differences among the treatments are not declared even though they actually exist.

Examples help clarify these two types of error. Consider the United States judicial system; the null hypothesis is "innocent until proven guilty beyond a reasonable shadow of a doubt" [10, p. 167]. The null hypothesis is stated this way purposely because our society is wary of the consequences of convicting innocent people: we therefore view "guiltiness" with some skepticism. Two correct decisions are possible (Fig. 2): (1) exonerating innocent defendants (accepting the null hypothesis when it is true), and (2) convicting guilty defendants (rejecting the null hypothesis when it is false). Convicting an innocent defendant is a Type 1 error, whereas exonerating a guilty one is a Type 2 error. This emphasizes the interrelatedness of the two error types. Because the evidence must be strong for juries to declare a defendant guilty, the Type 2 error rate is large. Similarly, in nursery experiments, if a researcher requires overwhelming evidence to reject the null hypothesis (i.e., declare differences among treatments), the Type 2 error rate will be high (i.e., some important treatment differences will be missed).

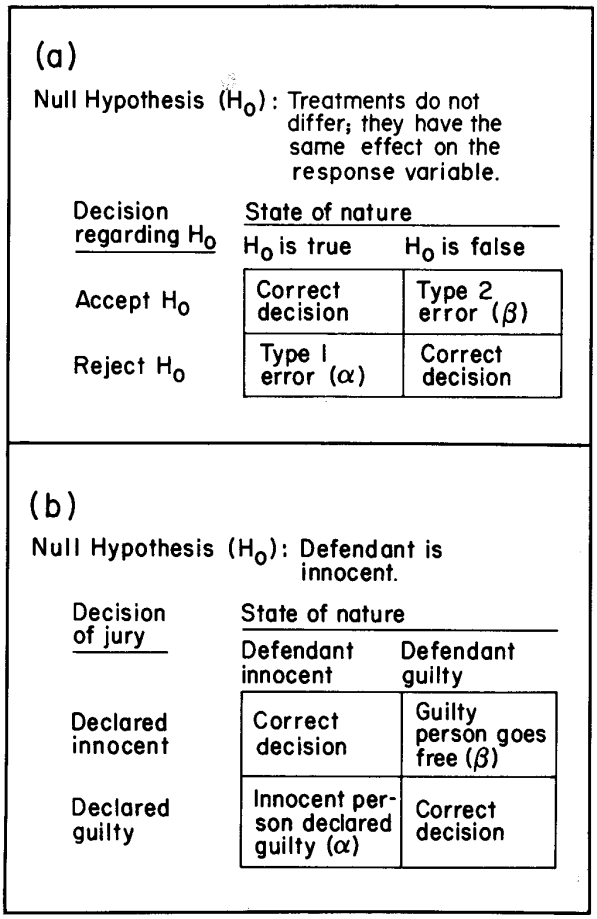

Figure 2. Correct and incorrect decisions are often made on the basis of incomplete information: (a) hypotheses tested by designed experiments and (b) a defendant judged by a jury (adapted from [10, p. 173]).

Only by increasing experimental precision can the rates of making both types of errors be reduced simultaneously. Increased precision is related to the notion of the **power,** or sensitivity, of the experiment. If Type 2 errors are infrequent (β is small), a researcher can be more confident of declaring treatment differences that actually exist, and the experiment is said to be powerful (sensitive to treatment differences).

In most investigations, the level of a is set by the experimenter; thus, the rate of Type 1 errors is known. Testing hypotheses at the $\alpha = 0.05$ level states explicitly that there is a 5% chance of declaring differences among treatments when they do not exist. On the other hand, β is often undetermined and in many cases extremely high. That is, biologically important differences among treatments are often missed (not declared significantly different) because the experiment is not powerful enough to detect them at $\alpha = 0.05$. The nursery researcher should always examine the magnitude of treatment differences and ask the biometrician for an approximation of the power of the experiment.

28.5.2 Replication and randomization

28.5.2.1 Replication

Replication is the repetition of treatments on more than one experimental plot [13, p. 5]. True replication means that a given treatment is applied independently to the multiple plots receiving that treatment. Because this last point causes considerable confusion in forestry experiments, it is useful to distinguish between *subsampling* and *replication.* For example, a nursery researcher applies two different fertilizer regimes to the entire length of each of two adjacent nursery beds. The researcher then scatters six 2-foot-long plots throughout each bed and counts the number of plantable seedlings lifted from each plot. In this instance, there are *not* six replicates of each treatment because the treatments are not applied independently to the six plots: in fact, all six plots received the same treatment application and are in the same bed. Rather, there is one replicate of each regime (a nursery bed) which is subsampled via subplots. To obtain true replication in the above example, 12 plots would first have to be chosen and the treatments then allocated to them at random (see 28.5.2.2).

The three functions of replication are to (1) increase the precision of estimated treatment effects, (2) provide a measure of experimental error, and (3) broaden the range of validity to

which the experimental conclusions apply. The first and third functions were recognized in agricultural experimentation as early as the 1700s, as farmers noticed that uncontrollable variation in yields from field to field and year to year made it impossible to recommend the use of one crop variety over another without comparing the two in a number of fields and years [5].

To illustrate these functions, suppose that the effects of two fertilizer regimes on 2+0 stem caliper are compared in adjacent plots in a number of different places (replicates) in the nursery. Plots receiving the same fertilizer treatment will vary in caliper due to uncontrolled variation in, say, soil fertility, soil texture, date sown, proximity to irrigation lines, and so on. When the fertilizers are compared over a large number of replicates, effects of these uncontrolled factors "average out," and the estimated difference between fertilizer regimes more precisely measures the true difference due to fertilizer. The uncontrolled variation among plots—the experimental error—can be measured by comparing replicates of the same treatment. If, for example, in replicate after replicate, regime 1 consistently results in larger caliper than regime 2, experimental error is small relative to treatment differences, and the researcher is likely to reject the null hypothesis that the two fertilizers affect caliper equally. Ideally, the experiment must be replicated under varying conditions of time and space. Repeating this same experiment over several years and nurseries extends the inference space (population of interest) in these two dimensions, broadening the range of validity of experimental results.

28.5.2.2 Randomization

"**Randomization** is the assignment of treatments to experimental plots so that all plots have an equal chance of receiving a treatment. It functions to assure unbiased estimates of treatment means and experimental error" [13, p. 5]. Put another way, randomization serves to equalize background (uncontrolled) characteristics of experimental plots receiving different treatments and provides a basis for statistical inference [1, p. 32]. Plots receiving one treatment should differ in no systematic way from plots receiving another treatment; this is accomplished in practice by drawing numbers out of a hat or by using random-number tables or random-number generators to match the treatments by chance to their assigned field plots.

Consider the following example to describe the reasons for randomization. Four replications of two fertilizer regimes are compared in the same nursery bed (Fig. 3). Two (of many) alternative field layouts include a systematic layout (Fig. 3a), in which regime 1 precedes regime 2 in each replicate, and a "random" layout (Fig. 3b). Treatment means (say, for caliper) are obtained by averaging the four replicates of each regime, and experimental error is estimated from the variation among plots receiving the same regime. Further suppose that a water gradient in this bed causes drainage to become consistently poorer from left to right. Because regime 1 always occurs before 2 in the systematic layout, it always experiences slightly more favorable drainage; this **bias** does *not* average out. Even if no true differences exist between the fertilizer regimes, mean seedling caliper (the response variable) may always be larger for regime 1 because it consistently experiences better drainage. Therefore, the estimated effect of regime 1 in the systematic layout is biased upwards from the confounding effects of water drainage.

only the extreme nature of the water gradient allowed recognition of the bias created by this particular systematic design. However, other systematic designs may suffer similar bias associated with gradients that researchers fail to recognize. Thus, random designs are essential as insurance against the possible bias generated by systematic variation in the uncontrolled characteristics of experimental plots.

Problems arise in experiments with few replications and treatments because "extreme" layouts—outcomes of randomization that appear systematic or unfavorable for some reason—occur fairly often, even at random. For example, if the experiment in Figure 3 were treated as a paired experiment with four pairs (see 28.5.3.1), the alternating scheme (12121212 or 21212121) would occur 1/8 of the time at random. Instances of such extreme layouts do not vitiate the need for randomization, but rather indicate the need for carefully examining the random layout before its field implementation. Cox [7, p. 86] presents an excellent discussion of methods for dealing with extreme outcomes.

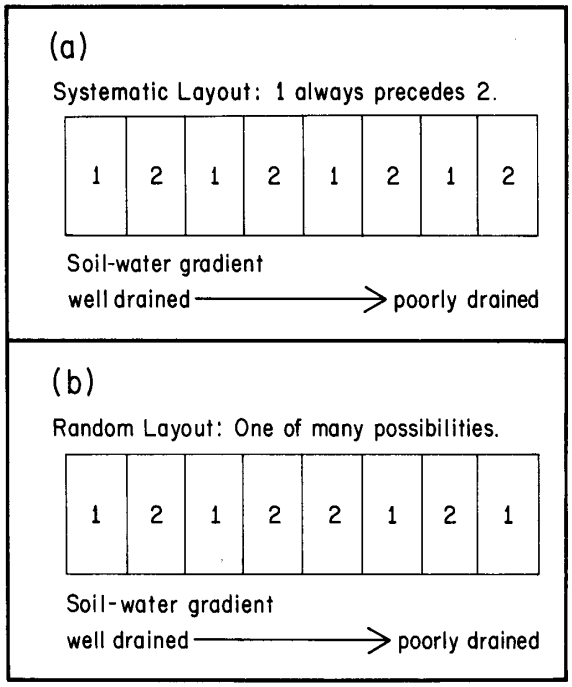

Figure 3. Two possible field layouts—(a) systematic and (b) random-of an experiment comprising four replications of two fertilizer regimes (1, 2) in a nursery bed.

28.5.3 Controlling experimental error

Reducing **experimental error** can greatly increase the sensitivity (power) of experiments to treatment differences. This section describes statistical methods of controlling error by choice of experimental design (see 28.5.3.1 to 28.5.3.3) and by covariance (see 28.5.3.4).

To this point, the discussion of experimental design suggests that treatments are always assigned to experimental plots totally at random. For example, in a nursery experiment with three replications of each of 10 treatments, randomization would ensure that each of the 30 nursery plots had an equal chance of receiving any replicate of any treatment. Such designs, called **completely randomized designs (CRDs),** are the simplest to lay out and analyze; however, experimental precision can often be increased by employing slightly more complex designs. Thus, imposing certain restrictions on the random assignment of treatments to experimental plots can control experimental error.

28.5.3.1 Pairing

The most intuitively appealing restriction of randomization occurs when an experiment testing the effects of two treatments is designed such that treatment assignments are made in pairs. Two similar experimental plots are identified and

called a pair; each plot within the pair randomly receives one of the two treatments. That is, randomization is still employed but is restricted to within-pair allocation of treatments to plots. The number of replicates equals the number of pairs. Because interest centers on comparing the relative, rather than absolute, effects of the two treatments, it is natural to use the *difference* between the paired treatment plots as the measure of response. Because experimental plots within pairs have similar background characteristics, uncontrolled variation (experimental error) is reduced by comparing differences on like plots.

An example should make clear these conceptual advantages of pairing. A nursery researcher, interested in comparing a new fertilizer regime to the standard regime for 2+0 Douglas-fir [*Pseudotsuga menziesii* (Mirb.) Franco], decides to test eight replications of the two treatments. To ensure broad validity of the results to the entire range of nursery conditions, the researcher first chooses eight nursery beds at random from all the beds containing Douglas-fir in their 2+0 season. Each of the 16 experimental plots will be 10 feet long and will receive one of the two regimes during the growing season; the number of plantable seedlings in the inner 4 feet of each plot (leaving a 3-foot border on each end) will be assessed at time of lifting. Because the two plots in each bed occupy only 20 feet, their position along the bed is also randomly located.

Two alternative designs for this experiment are shown in Figure 4. For the CRD (Fig. 4a), each of the 16 plots had an equal chance of receiving one of the two treatment regimes; one (of many) possible schemes is shown. For the paired-plot arrangement, one regime was randomly assigned to either of the two plots per bed by a coin flip, the other to the remaining member of the pair.

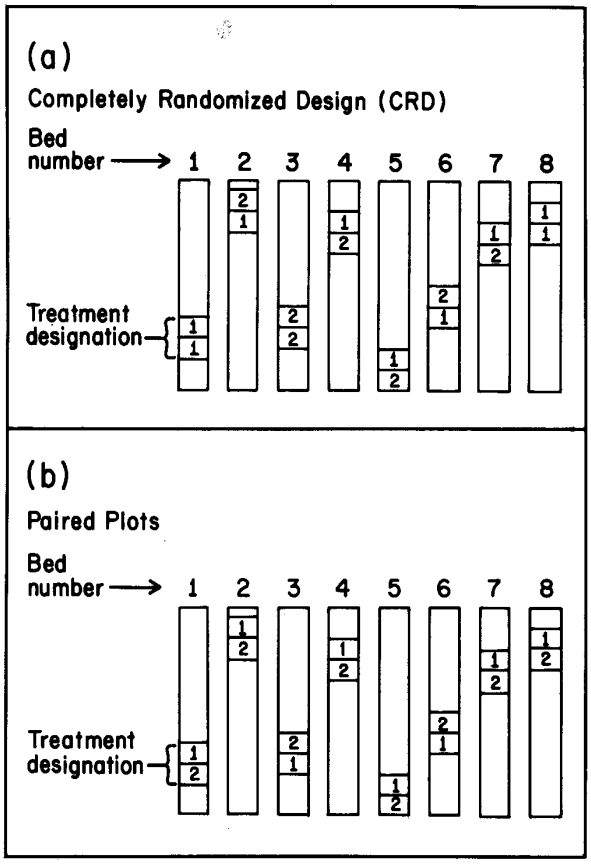

Figure 4. Two alternative experimental designs testing eight replications of two fertilizer regimes (1, 2). The field layouts of nursery beds and the plot locations within beds are chosen at random. Fertilizer regimes are assigned (a) at random to the 16 plots and (b) to the 16 plots as eight pairs, each member of the pair receiving one of the two fertilizer regimes.

Imagine that these 16 plots sample a wide range of drainage conditions, fertility, soil texture, and proximity to irrigation lines. Adjacent plots in the same bed will be more similar with respect to these conditions than will plots in different beds. Pairing plots exploits this similarity by basing the analysis of treatment effects on the difference between treatments occurring in the same bed. Two plots lying in a poorly drained area of a bed necessarily have reduced yield of plantable seedlings. In the CRD, one fertilizer regime may be assigned to both plots of the poorly drained bed location, reducing its average over the eight replicates and increasing experimental error. In the paired-plot arrangement, however, the two regimes will be negatively affected in the same way on a poorly drained location; the differential effect of regime 1 over regime 2 may remain relatively stable, except for other sources of background variation associated with differences between plots within pairs.

28.5.3.2 Blocking

The concept of pairing logically extends to that of blocking for experiments with more than two treatments. In fact, paired plots are the simplest case of blocking. Suppose, for an experiment testing a new herbicide at two application rates (L = low, H = high) against both a control (C = no herbicide) and the standard herbicide (S = standard), that each of the four treatments is replicated 5 times (20 plots). If a CRD is used (Fig. 5a), treatments are assigned totally at random. If a **randomized complete block (RCB)** is used (Fig. 5b), each nursery bed is assigned one complete replicate of the experiment; randomization is restricted to the allocation of treatments *within a block.* If nursery beds are scattered, representing a wide variety of conditions, plots within a bed should be more similar to one another than to those from different beds.

Consider what this does to comparisons of treatment effects for the above-mentioned herbicide experiment. At the end of the first growing season, height of 100 seedlings within each plot is measured to determine the possible phytotoxic effects of the new chemical. Suppose that bed 5 was not well prepared by the bed former and that this, combined with the inherent soil attributes of that bed, reduces height growth in bed 5 regardless of treatment. Bed 4, on the other hand, is located in a part of the nursery recently mulched, which accelerates seedling growth. These "bed effects" are average effects on 1+0 height common to all plots in a given bed. In addition to bed and treatment effects, 1+0 height also is influenced by "plot effects"—uncontrolled variation due to background characteristics of the specific plots within beds. For any of the 20 plots, then, 1+0 height may be expressed as the sum of these three effects:

Height = (treatment effect) + (bed effect) + (plot effect)

The mean for each of the four treatments is estimated as its average over five replicates. In the case of the RCB, bed effects influence each treatment mean equally because each treatment occurs once in each bed. in the CRD, differential bed effects influence treatment averages; specifically, the negative effect of bed 5 reduces the average levels for L and H of the new chemical, whereas the positive effect of bed 4 increases the average height in two plots for S. Thus, the RCB design increases the precision with which treatment effects are estimated by allowing bed effects to be estimated separately and removed from the comparisons of treatments.

This increase in precision is reflected in reduced experimental error. Overall error is estimated by the variability among the five plots receiving the same treatment. In the CRD, the five plots for any one treatment vary both by bed and plot effects,

and both contribute to experimental error. In the RCB, bed effects can be directly estimated and can be eliminated from the experimental error.

It is critical to realize conceptually that blocking works any time experimental plots of an entire replication (one plot of each treatment) can be grouped such that they are more similar to one another than to plots of other blocks. Then, differences among the groups are termed block effects (bed effects, in the herbicide experiment). This may mean, for example, that blocks are sown on different days, measured by different observers, or located in different parts of the nursery (see 28.6 for practical implications of blocking).

Finally, the concept of blocking can be extended to more than one dimension through Latin Squares and similar designs. Though sometimes useful in nursery research, these designs suffer from sensitivity to missing data (e.g., plots accidentally destroyed by faulty irrigation) and from restrictions on the number of treatments and replications. Neter and Wasserman [15] present an excellent discussion on design and analysis and on overcoming problems of these designs.

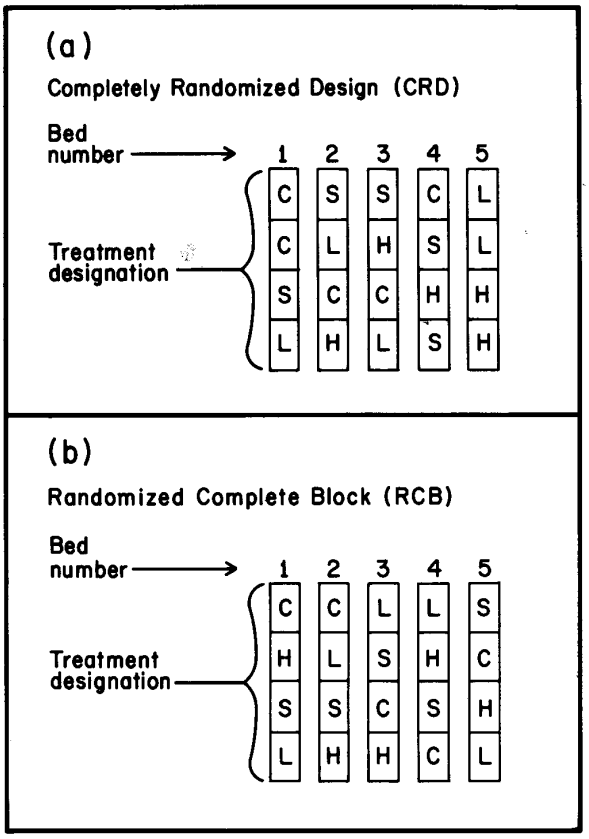

Figure 5. Two alternative experimental designs testing four herbicide treatments (C = control, S = standard herbicide, L = low level of new herbicide, H = high level of new herbicide). The benefit of the RCB (b) over the CRD (a) is that each bed receives one replication of all four treatments.

28.5.3.3 Split-plot principle

When, for practical reasons, some factors of an experiment require larger plot sizes than others, the split-plot principle is often applied. For instance, the minimum size of plots for irrigation treatments and sowing dates is necessarily larger than that for different sowing densities and seed sources. A split-plot design for two factors calls for assigning treatments of one factor to larger plots (called main plots or whole plots) in a CRD, RCB, or other design and then splitting each whole plot into enough subplots to accommodate one replicate of each treatment level of the second factor. Because each whole plot contains a complete replication of the treatments of the second factor, it is a "block" of the second factor. Randomization occurs in two stages: first, in assigning treatments of factor 1 to whole plots, then, in assigning treatments of factor 2 to subplots within each whole plot. Precision is often sacrificed for estimating effects of the whole-plot factor, but increased for subplot treatment comparisons. Split-plot designs can be extended to multiple factors at both the whole-plot and subplot level and even to splitting the subplot (split-split plots). Cox [7, p. 142] and Little and Hills [13, chapters 8, 9] give excellent accounts of the concepts and algebra of split-plot designs; Cochran and Cox [6, chapter 7] present a more advanced treatment.

For the purposes of describing the concepts behind split-plot designs, consider a two-factor nursery experiment investigating the effects of three fertilization regimes (F1, F2, F3) at each of two irrigation levels (H = high, L = low) on stem caliper of 2+1 Douglas-fir (Fig. 6). The irrigation system in the nursery may require that several beds on either side of an irrigation line receive the same irrigation treatment. It may be that the nursery researcher can only devote six lines to the entire experiment (say, two lines in each of three different sections of the nursery). Therefore, a possible RCB design for irrigation (exclusive of fertilization) may be obtained by randomly assigning one of the two irrigation treatments to one of the two lines within each section (block) (Fig. 6a); this is an RCB with two treatments and three blocks. All beds watered by each line receive the same irrigation treatment. The fertilizer treatment may then be added by "splitting" each bed into three lengths (subplots) to which one of the three fertilizer regimes is randomly assigned (Fig. 6b). Thus, the randomization of fertilizer treatments is restricted to allocation within an irrigation whole-plot.

Regardless of the response variable, two types of experimental error are associated with this experiment. The subplot error, resulting from residual variation among subplots, estimates microsite and other background differences influencing the response of subplots within a whole plot. The whole-plot error, resulting from the uncontrolled background variation among whole plots within a block, is usually larger than the subplot error.

28.5.3.4 Covariance

In a previous section (28.5.3.2), blocking of experimental plots into groups of similar soil types, etc. was presented as a method of reducing experimental error. Even after plots are grouped, however, background characteristics of plots within a block may still vary. Knowing that this variation exists may be used to reduce experimental error by the statistical process of covariance [7, chapter 4; 18, chapter 15]. Covariance requires making additional measurements of these other characteristics (called **concomitant variables)** on each experimental plot.

For an experiment testing the effects of different types of root wrenching on seedling caliper, suppose that even after blocking, substantial variation in soil N level exists among plots within blocks. N level may have an average influence on caliper; higher inherent N means larger average caliper. Knowing this relationship may help the researcher adjust treatment means to a common starting value of soil N.

The analysis of covariance may be used for any type of statistical design (e.g., CRD, RCB, Latin Square) as an additional method of increasing precision. Foresters routinely use the simplest form of covariance by analyzing height "growth" instead of total height as a measure of treatment response. The rationale is to remove some of the initial variation in

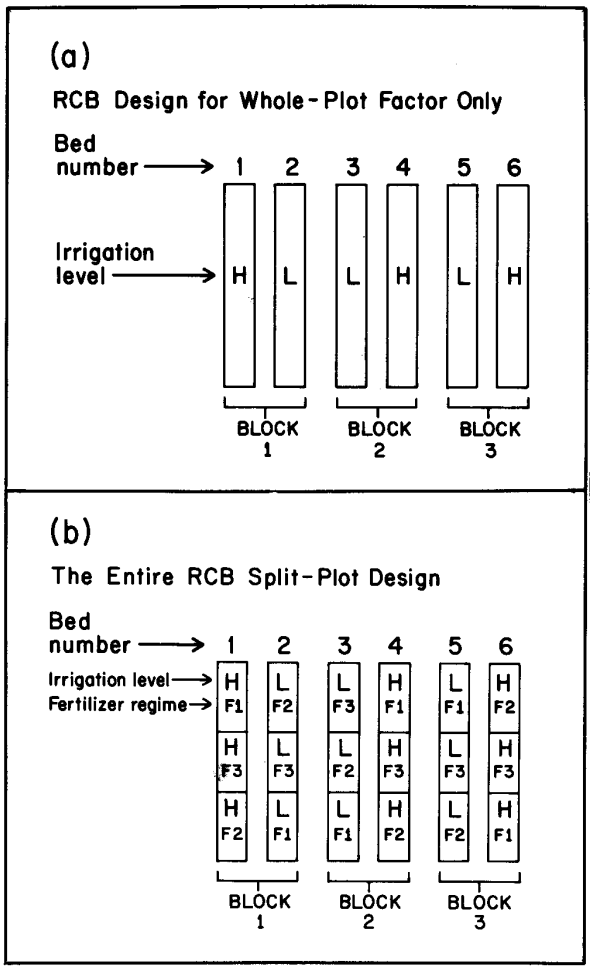

Figure 6. RCB split-plot design testing two irrigation levels (H = high, L = low) and three fertilizer regimes (F1, F2, F3): (a) irrigation, requiring larger plots, is the whole-plot factor; (b) fertilization, allowing smaller plots, is the subplot factor.

height by measuring height before and after treatment application and analyzing just the portion of height (the increment) added after treatment.

Some words of caution are in order, however. The interpretation of covariance analysis can become very complicated, making it difficult to separate the effects of the concomitant variable from those of the treatments. Statistically, this happens for complex experimental designs and when some plots are missing. Biologically, this happens when the concomitant variable is affected by the treatment [7, p. 48]. This last point is extremely important. For example, if soil P level is used as a concomitant variable in an experiment assessing effects of phosphate fertilization on caliper and if it is measured *after* fertilizer application, then the treatment will drastically influence the concomitant variable. Adjusting for the average effect of soil P on caliper might eliminate all treatment (fertilizer) effects. To avoid this type of problem, make sure that concomitant variables are (1) observed before treatments are applied or (2) unaffected by the treatments.

28.6 The Research Process for Designed Nursery Experiments

This section covers in some depth the research process (28.3) applied to designed nursery experiments. Emphasis is placed on the importance of statistical concepts (28.5) in developing, designing, conducting, analyzing, and interpreting these experiments.

The following discussion provides a checklist for managers as they encounter and attack problems in their nurseries. Though presented in chronological order, the steps of the research process are not independent of each other. Often, understanding the rationale of subsequent steps helps accomplish the current one.

28.6.1 Defining the problem

28.6.1.1 Recognizing a problem

Most nursery experiments stem from problems requiring practical solutions (see chapter 29, this volume). Even nursery research investigating the fundamental principles underlying a problem-such as the physiological basis of increased outplanting vigor after fall fertilization of 2+0s—has immediate practical application. Practical problems may be arbitrarily classified as either today's limitations or tomorrow's opportunities. Today's limitations include existing insect, pathogen, or drainage problems, or optimizing packing-line arrangement with existing equipment; tomorrow's opportunities involve incorporating new technology—whether new chemicals, machinery, or cultural practices—to improve seedling quality and cost-effectiveness. Nursery researchers must be adept at recognizing and addressing both problem types.

28.6.1.2 Refining the problem

Once a problem area has been identified, the nursery researcher relies on personal experience and secures existing information from literature, chemical labels, other nursery personnel, and specialists (e.g., extension agent) to answer the following questions:

Does the problem warrant research? The answer may be no if (1) immediate action is required, (2) results from other studies are conclusive and broadly applicable, (3) cost or effort involved in doing the research is high relative to potential benefits, or (4) cost of implementing the results is too high.

If the problem is researchable, what specific questions remain to be answered? Perhaps some parts of the problem seem solved, but others need further investigation. For example, a new herbicide may have been demonstrated safe and effective in several nurseries, but optimal application rates and timing remain to be determined for your nursery conditions.

is a designed experiment needed? Perhaps the attributes of designed experiments (see 28.4.2) are *not* required and an operational trial (see 28.4.1) will answer the questions for less cost and effort.

To what population should the results apply? The inference space [2, p. 84] or range of validity of the results must be defined. For instance, should the results and conclusions apply to (1) 1+0 Douglas-fir in one particular nursery block, (2) 1+0 Douglas-fir in poorly drained parts of the nursery, (3) 1+0 Douglas-fir in all bareroot nurseries west of the Cascade Mountains, or (4) 1+0 and 2+0 conifers in all bareroot nurseries west of the Cascades?

28.6.2 Setting objectives

Delineating the experimental objectives serves to clearly state the problems to be addressed and sets the framework for specifying experimental methodology [3, p. 38]. Objectives can take several forms: (1) to determine the effects of a certain factor, for example, of fall fertilization with N; (2) to investigate interactions, for example, of irrigation and fertilization; (3) to

find optimum application rates, for example, of herbicides or fertilizers; and (4) to develop prediction equations, for example, of the average shoot:root ratio of a batch of seedlings, given their average height, caliper, and seed source. Although statistical null hypotheses are often implied when research objectives are stated, their explicit delineation is left until the experiment has been designed (28.6.3.5).

Objectives should be ranked to ensure that the experiment is designed to answer the most important questions first. Often, it is possible (for example, by split-plot design) to increase the precision with which certain treatment means are estimated or hypotheses tested by sacrificing precision on others.

28.6.3 Planning the experiment

The nursery researcher frequently directs the early efforts of planning an experiment (see 28.6.3.1 to 28.6.3.3), whereas the biometrician may direct the later ones (see 28.6.3.4 to 28.6.3.7). Though the leader may change, a true team effort is required throughout to ensure that the experiment meets its stated objectives. For instance, the nursery researcher may list factors to investigate and the biometrician then help determine the levels at which to test each factor. Conversely, the biometrician may set a minimum number of replications needed and the nursery researcher put an upper practical limit on the number feasible. A constant balance is needed between what might appear statistically favorable and what is practically suited to experimental conditions in the nursery.

28.6.3.1 Choosing treatments

Controls and standards.—Nursery experiments most often require a treatment that serves as the basis against which the effectiveness of other treatments is judged. Controls and standards both serve this purpose.

A **control treatment** is the zero-level application of a factor; no wrenching, 0 pounds of N fertilizer, no irrigation, and no herbicide are examples of controls that might be used to judge whether particular treatment levels of wrenching, N fertilization, irrigation, and herbicide spray, respectively, were effective. Control treatments are particularly useful when an unproven or new factor is being tested but are less so when the experimental objective is to find an optimum level of a "known-effective" factor.

Standard treatments are the "standard operating procedures." A new treatment must often prove itself against the standard to justify altering current practices. For instance, a new alignment of personnel in the packing shed must be proven superior to the current one to justify switching.

Single-factor versus multifactor experiments.—Single-factor experiments, in which only one condition (factor) is varied among the treatments, are commonly used in nurseries at either end of the research process: operational trials or final-stage experiments. In operational trials, the researcher might test three new herbicides for control of grasses or compare two seeders for evenness of sowing; in both cases, only one factor, herbicide or equipment type, is varied. In the final stages of experimentation, the researcher often knows the proper levels at which to control nontreatment conditions and varies only the critical factor of interest (say, herbicide application rate) to find the optimum level [6, p. 152].

Most nursery experiments lie between these two extremes, exploring the effects of one factor (say, bed density) over various levels of other factors (sowing dates and species). Such multifactor experiments often test, for example, whether the most effective bed density is the same for all sowing dates and species.

Factorial experiments.—Factorials are by far the most common arrangement of treatments in multifactor experiments. Factorial experiments test each level of each factor at all levels of the other factors. In a three-factor experiment with two bed densities, three N levels, and two seed zones of Douglas-fir, there are 2 x 3 x 2 = 12 treatments. Each treatment consists of a specified level of each of the three factors—for example, treatment 1 might be low bed density, no N (the control), and seed zone 062; 12 separate treatments are required to test each factor across all levels of the other factors. These 12 treatments can be applied to experimental plots in a variety of experimental designs (CRD, RCB, or Latin Square); "factorial" just defines the number and structure of the treatments, not the field design.

The nature of factorials and the reasons for their importance are discussed fully in Cox [7, p. 94]. Briefly, factorials allow explicit investigation of the interaction among factors. If interactions are not significant, then factorial experiments extend the range of validity and increase the precision of estimating factor effects, relative to separate experiments of the individual factors. For example, the effects of N and P may be investigated either in separate experiments or in one factorial experiment. If experiments are done separately, N is held at a constant (standard) level while multiple P levels are investigated; conversely, P is held at a constant level while multiple N levels are investigated. The factorial allows elucidation of interactions because rates of N and P are varied together so that all combinations of both factors are applied; for example, N may increase caliper only in the presence of high P levels. In the absence of interactions (that is, if the effects of N do *not* depend on the level of P, and vice versa), the range of validity is extended because the researcher knows that each nutrient is effective over several levels of the other, *not* just the standard. The precision of estimating effects is also increased [7, p. 94].

Factorials are not without their drawbacks [6, p. 152]. But for most nursery experiments, these are more a matter of the complexity of the problem than the factorial itself. Factorials can often become large (for example, a 3 x 3 x 5 factorial has 45 treatments), making them difficult to implement and, sometimes, interpret; however, the efficiency of factorial arrangements, compared to that of separate experiments, increases for large, multifactor experiments.

Choosing factors.—While mainly directed to factorials, the discussion here applies rather broadly to choosing factors in multifactor experiments [7, p. 134]. For the most part, multifactor experiments examine only one or two factors of primary interest; these **primary factors** are the reason for the experiment. **Supplementary or subsidiary** factors [6, p. 151] may be added to (1) shed light on the mode of action, (2) extend the range of validity, and (or) (3) determine interactions with the primary factor(s).

In an experiment testing the effects of fall N fertilization on the frost hardiness of 2+0 Douglas-fir, the primary factor is N; supplementary factors might include seed zone and irrigation. The N levels are tried (1) at various irrigation levels, both to examine interactions (perhaps standard irrigation is not best with fall fertilization) and to elucidate N's mode of action in increasing frost hardiness (perhaps moister conditions promote the physiological actions of N relating to frost hardiness), and (2) at different seed zones, to provide a wider range of validity if N acts consistently across all the zones tested. The nursery researcher should choose primary factors that meet the experimental objectives and supplementary factors that ensure general conclusions can be drawn about the primary factors over the intended population.

Choosing factor levels.—In choosing levels at which to test each factor, the nursery researcher must again consider the experimental objectives and intended population. For quantitative factors whose levels represent points on a continuum (e.g., pounds of fertilizer), the range of levels should bracket the range expected to be operationally feasible [7, p. 141]. For example, the lowest level of fertilization or chemical application is often zero-the control: the upper level is chosen on the basis of experience, cost, or other information as the upper practical extreme. How many levels and where to position the levels depend on the nature of the response curve (linear, quadratic, asymptotic) and the purpose of the experiment (see [7, p. 141] for a full discussion). For many nursery experiments, three or four well-spaced levels (including a control, if appropriate) are sufficient.

For supplementary factors included to extend the range of validity or to detect interactions, a few extreme levels often suffice. For example, if N fertilizer causes similar responses in extreme seed sources of coastal Douglas-fir—say, west Cascade, valley, and coastal—it may be safe to extend the experimental conclusions to all coastal Douglas-fir. And if N shows no interaction with P over a wide range of levels, then inferences over the entire range are valid. The only caution here is to use factor levels of interest. If Cascades Douglas-fir is not grown at the nursery, then why include it unless more fundamental questions about the Douglas-fir species are of interest. Again, keep the population of interest in mind when choosing both factors and their levels.

Choosing levels for nontreatment conditions.—Once the treatments have been determined, it is critical that the conditions or factors *not* varied are held constant at meaningful levels. For example, in an experiment testing the relative effectiveness of two root-wrenching depths, only depth is varied. Other factors—such as seed zone, stratification period, sowing date, and fertilization and irrigation regimes—are held constant. In this instance, the rate of irrigation may have a dramatic impact on treatment effectiveness; thus, its level (though not varied) is critical to interpreting the results. Often, but not always, the nontreatment factors are held constant at their normal or standard levels to test the effectiveness of the treatments if everything else is done as usual.

28.6.3.2 Choosing variables to measure

The nursery researcher is faced with a wide array of variables that could be affected by or could affect treatment responses. Which variables to measure depends on how much time and effort are available and how likely it is that a particular variable may be of practical value. Variables measured fall into three broad categories; keeping these categories in mind can often help researchers decide which variables are pertinent.

Response variables.—Response variables are those that the treatments were meant to test. They are usually measured on all **observational units** (items to be experimentally measured or observed; for example, seedlings) in a **measurement plot** (portion of the experimental plot actually measured), then aggregated to obtain a plot mean or sum. The most critical of these are usually delineated directly in the experimental objectives. Height, caliper, shoot:root ratio, number of plantable seedlings, foliar N levels, frost hardiness, and outplanting growth and survival are examples of often-measured nursery response variables.

Response variables may be either quantitative (numeric) or categorical (falling into discrete classes). The numeric are routinely analyzed with a technique termed analysis of variance, the categorical by other techniques [17] or by assigning numbers to the classes (categories) to make them pseudonumeric. For example, six vigor classes might arbitrarily be assigned the numbers 1 (low) to 6 (high). Caution must be used here because this implies that class 6 is 1 unit better than class 5, 2 units better than class 4, and so on. The biological basis of such assignments should be weighed. For categorical data, more classes mean more discrimination among treatments unless responses *cannot* be accurately assigned. Four and six classes are often useful numbers of classes for assigning biological responses; an even number is recommended because of the psychological tendency of observers to overassign responses to a middle category (such as to the third class, if five classes were available).

Concomitant variables.—Concomitant variables are those measured on each experimental plot or observational unit (seedling), for the purposes of using covariance analysis (see 28.5.3.4). The precision of the experiment can be increased by adjusting response variables to a common, average level of the concomitant variable. Concomitant variables—for example, pretreatment soil N levels, initial bed density, or soil textures—must be measured on each experimental plot, be independent of treatment effects, and be numeric.

Explanatory variables.—Explanatory variables are often measured to shed light on underlying principles of action or to document experimental conditions. These variables can be measured at any level. On the experimental-plot level, the nursery researcher might test to see whether fertilizer or chemicals were applied properly by assaying each plot shortly after application. On the block level (if beds are blocks), the researcher might monitor plant water potential at various points in nursery beds situated varying distances from the irrigation lines; large differences in seedling growth from block to block may then be related to water status. Finally, on the nursery level, the researcher might monitor climatic conditions relative to sowing date; early sowing may pay handsome dividends in some years, whereas its effects may be disastrous during other years with different spring weather.

In general, explanatory variables are measured for biological or physical, *not* statistical, reasons. They are often used in the deductive process to extrapolate or "explain" experimental results.

28.6.3.3 Determining plot size

Determining the appropriate size and shape of experimental plots requires both statistical and practical considerations. For a specified amount of land devoted to an experiment, the number of replications necessarily decreases as plot size increases. As a rule, once a *minimum* plot size is reached, precision is increased more effectively by adding replications than by enlarging plots [9, p. 3]. Practical considerations, subsequently described, often loom large in determining this minimum plot size.

Remember that the experimental plot (say, a length of nursery bed) is the smallest physical unit to which a treatment is applied independent of all other treatments. All observational units (seedlings) within a plot do *not* have to be measured. It is largely the responsibility of the nursery manager to ensure that the total size of the experimental plot satisfies practical constraints and of the researcher and biometrician together to determine the size of the measurement plot within each experimental plot.

Plot shape, usually constrained by bed shape in forest-tree nurseries, will not be addressed here; references include Le Clerg et al. [12] and Gomez [9].

Experimental plots.—Practical considerations influencing size of the experimental plot fall into three overlapping categories: operational constraints, representation, and independence.

- **Operational constraints:** These refer to physical limitations in the ability to apply treatments independently to **small plots.** Such limits often lead to split-plot designs (see 28.5.3.3), in which one factor (the whole-plot factor) is applied to a much larger plot than other factors (subplot factors) in the experiment. Irrigation is a good case in point; the physical nature of the system often requires that several beds on either side of a line *must* receive the same irrigation level. Thus, a group of beds is the smallest physical unit to which the researcher can independently (randomly) assign different irrigation levels. Other treatments (the subplot treatments) in the split-plot design, such as fertilizer levels or different wrenching techniques, can be randomly assigned to smaller plots within this group of beds.

  For multifactor experiments, the researcher should construct a brief list detailing the operationally feasible plot sizes for the factors under investigation. This list also is useful for determining experimental designs and conducting the experiment.

- **Representation:** In most nursery experiments, it is critical that the conditions imposed by the experimental treatments be representative of those same treatments applied operationally [7, p. 194]. For instance, the artificial nature of irrigating by hand may be intolerable even though it allows smaller plot sizes. Many treatments applied in nursery experiments have start-stop problems in the sense that representative treatments are *not* attained at the beginning and ending of each plot; for example, many seeders disperse seed unevenly for the first and last foot or two. Thus, the total plot size must be large enough to leave a representative measurement plot in the "middle," the "ends" serving as borders.

  Representation, as used here, relates to bias and inaccuracy that can result from nonrepresentative plots. Careful, precise experimental technique *cannot* overcome bias resulting in this manner. For example, in time and motion studies such as might be conducted to investigate alternative packing-line arrangements, the experimental time allotted must be long enough to accurately represent the operational situation. Some arrangements, faster in the short run, may cause workers to take more breaks or to suffer more illness or boredom when imposed under normal, operating conditions.

  Representation is less important in fundamental studies investigating the basic biological principles underlying treatment response. There, statistical precision and choice of treatments to illuminate reasons for response are most important; hand application of fertilizer or irrigation and manual sowing (or thinning to desired bed densities) may be entirely suitable.

- **Independence:** Treatment application and (or) response on a particular plot should *not* influence response on adjacent plots [7, p. 196]. For example, spread of sprays, fertilizer, and water can unknowingly affect growth on near-by plots. In time and motion studies, two packing-line arrangements tried on successive dates might allow a carry-over effect from the first day's arrangement to the second. Researchers should make every effort to ensure that experimental plots are large enough for measurement plots to respond independently.

Measurement plots.—In most field experiments, the actual measurement plot is a subplot nestled within the experimental plot. The unmeasured observational units (seedlings) surrounding the measurement plot buffer that plot from edge effects and from effects caused by treatments on other experimental plots.

In addition to independence and representation, the number of trees in a measurement plot and its orientation and location are the major concerns.

- **Edge effects:** These occur both on the ends and sides of nursery-bed plots. As a very general rule, end effects are bad (artifacts of the experiment), and side effects are good (representative of the nursery).

  End effects usually occur as a result of stop-start problems associated with treatment application to small experimental plots (see **Representation,** just discussed). They should be avoided by placing measurement plots away from the ends of experimental plots; these ends should be left to border the measurement plot, buffering it from the external influences of adjacent treatments and making it more nearly like a randomly chosen location in the middle of an operational bed.

  Side effects occur because the outermost drill row along each side of the nursery bed tends to grow and respond to treatment differently than the inner rows. However, this type of edge effect would occur to the side rows if the treatment were applied on an operational scale and in practical nursery experiments. Because the inferences drawn should pertain to the population of *all* seedlings, these outermost rows should be included in the measurement plot to make it as representative as possible of the population of interest. Thus, a measurement plot for practical experiments should be a swath that stretches across the entire seedbed; each row contributes equally to the plot mean just as each row contributes equally to the harvestable crop.

- **Response variable:** Depending upon the trait being measured, fewer or more trees need be included in the measurement plot to obtain a precise plot mean for the treatment. The amount of effort and expense required to obtain each measurement also influences the number included. As a general rule, for traits like height, caliper, and number of plantable seedlings, a 1-foot section of bed provides a more than adequate number of seedlings (~ 100 at 25/ft^2) and is easy to lay out.

- **Subsampling:** Multiple measurement plots are often placed within one experimental plot either to allow for multiple destructive sampling throughout the growing season (e.g., when shoot:root ratio is assessed at multiple times during the growing season) or to provide an estimate of within-plot variability. In the case of such subsampling,.the following considerations apply to each measurement plot: (1) handling of one measurement plot should not influence response on adjacent ones, often necessitating that buffer areas be left between measurement plots; (2) each measurement plot should represent the population of interest; and (3) enough seedlings should be included in each measurement plot to provide a precise plot mean.

28.6.3.4 Choosing an experimental design

By this point, tentative decisions have been made (mostly by the nursery researcher); regarding factors and factor levels to investigate, variables to measure, and practical limits on experimental- and measurement-plot sizes. The biometrician and researcher now employ the concepts of randomization, replication, and error control (see 28.5) to develop a statistically and operationally appropriate experimental design.

Applying these concepts to nursery field experiments produces frequent use of only a few common designs. CRDs (see Figs. 4a and 5a), in which the random assignment of treatments to plots is unrestricted, are *not* common in nursery

experimentation. A logical field grouping of experimental plots almost always exists such that plots within a group are more similar to each other with respect to water drainage, proximity to irrigation, and (or) soil characteristics than are plots from different groups. As a result, the RCB design (see Figs. 5b and 6) is the most commonly used in nurseries. In addition, RCBs are relatively easy to lay out and analyze and are relatively insensitive to the accidental loss or destruction of a plot or two. Latin Squares are used, though less frequently, when bidirectional field gradients exist; however, they suffer from the drawbacks discussed in section 28.5.3.2.

For multifactor experiments, especially those investigating irrigation, sowing date, or wrenching, split-plot designs are common (see Fig. 6). These most often have the whole-plot factor (such as irrigation) arranged in randomized complete blocks, with the treatments of one or two subplot factors (such as three levels of fertilization and two bed densities) arranged totally at random within each whole plot. Although these designs are common, more complex ones are warranted for large experiments that "mushroom" and would occupy too much space and require too much effort if many replications of each treatment were applied. The biometrician and researcher must confer and be innovative to arrive at appropriate designs (such as fractional factorials) for these more complex experimental situations.

28.6.3.5 Determining the number of replications

Theoretical considerations.—Many factors (practical and statistical) impinge on the number of replications required for a particular experiment. The effects of these factors are described here, both mathematically and intuitively, but understanding their effects does not depend on the mathematical relationships; it is added only for completeness. The practical nursery implications for each factor are delineated.

For RCB designs, the factors influencing the number of replications (i.e., blocks) interact through the formula

$$n = \frac{4t_\alpha^2 (CV)^2}{D^2} \qquad (1)$$

where n = number of blocks
D = detectable level of difference (%) between two treatments
t_α = tabular value of t for a specified Type 1 error rate (α) and number of degrees of freedom
CV = coefficient of variation (%) obtained as (mean square error)$^{1/2}$/experimental mean.

The **detectable level of difference** (D) is that difference between two treatment means (expressed as a percentage) which the experiment is able to declare significant. For example, if height of 2+0 seedlings increases from 15 to 18 inches as a result of fertilization, D = 20% (a 20% increase). In general, smaller differences are more difficult to detect (declare statistically significant) and require more replications. The researcher should decide in advance, roughly, the differences among treatments that represent biologically or practically important responses.

Recall that the **Type 1 error rate** (α) is the probability of declaring treatments significantly different when, in fact, they are not. The nursery researcher will necessarily set a low Type 1 error rate if a high degree of confidence is required before drawing conclusions from an experiment. The higher level of confidence required necessitates more replications to declare results significant. For example, for a given level of detection (say, D = 20%), more replications are required to declare results significant at $\alpha = 0.01$ (99% confidence level) than at $\alpha = 0.05$ (95% confidence level) (Fig. 7).

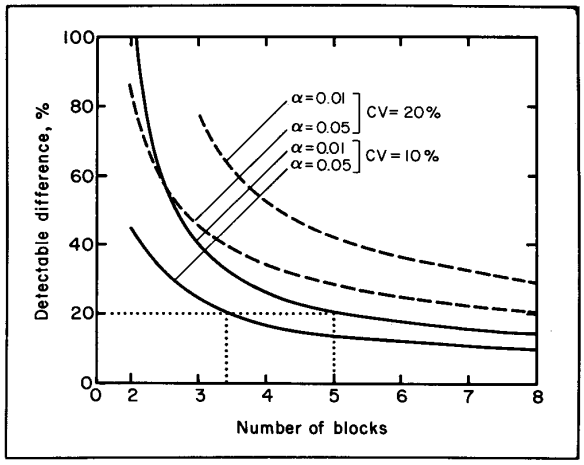

Figure 7. Effects of the number of blocks on the level of detectable difference (D) among treatments for two levels of inherent variability (coefficient of variation, CV, = 10 and 20%) and two levels of Type 1 error rate (a = 0.05 and 0.01) for an RCB with three treatments. Dotted line Indicates the number of blocks required at two a levels when D = 20%.

In principle, the **coefficient of variation** (CV) measures the background variation among plots receiving the same treatment as a percentage of the treatment mean. Thus, for a given level of detection, more replications are required for traits with higher CVs because the higher variability means larger experimental error (Fig. 7). For nursery experiments, CVs are influenced by (1) the response variable (e.g., root-growth capacity is extremely variable), (2) the experimental material (1+0 heights are more variable on a percentage basis than 2+0 heights), and (3) the variability among field plots. CVs between 10 and 20% are common in nursery field experiments.

In general, experiments with more treatments require fewer replications. In rough terms, each treatment provides an estimate of experimental error via the variation among the experimental plots receiving that treatment; these are pooled (combined) to give an average "experiment-wide" estimate of error. More treatments result in more of these individual estimates and thus a more precise estimate of the experimental error. When both the replications and the number of treatments are

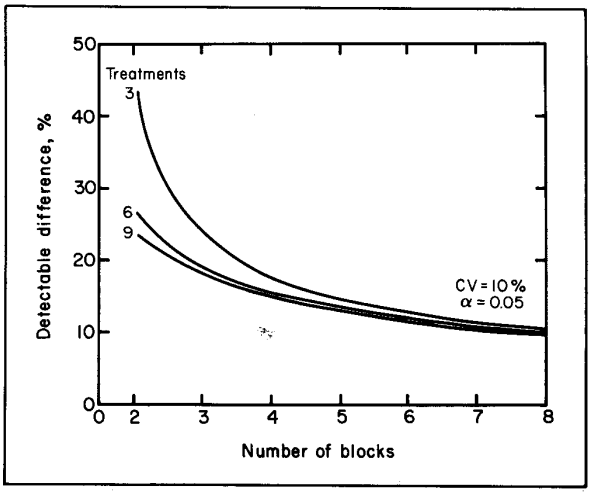

Figure 8. Effects of the number of blocks on the level of detectable difference between two treatment means for an RCB with different numbers of treatments (3, 6, 9).

small, an increase in either is especially effective in increasing the sensitivity of the experiment to detecting differences (Fig. 8).

Appropriate experimental designs can effectively reduce the number of replications needed for a given desired level of detection by eliminating extraneous sources of experimental error from treatment comparisons. This increases precision by reducing the experimental error and thus the CV. Furthermore, subsampling within an experimental plot can sometimes reduce the number of replications needed by providing an additional estimate of experimental error. The use of subsampling depends on many factors [4] but is most beneficial when the number of replications is limited for practical reasons.

Practical recommendations.—In addition to the previously stated statistical considerations, the number of replications required is influenced greatly by several practical factors. If the chance of incurring missing plots is high, more replications are needed. In addition, the accuracy of the experimental work is extremely critical in reducing the CV and thus lowering the number of replications required.

In general, three replications are a minimum, and more are required if (1) the number of treatments is small, (2) the response variable has a high CV, or (3) small differences should be detectable with high confidence. When estimates of the factors in Equation I are available, the number of blocks (for an RCB) can be calculated. When data are not available, a rule of thumb [14] is to choose the number of blocks, n, such that $(n - 1)(t - 1) \geq 12$, where t is the number of treatments; this ensures 12 degrees of freedom for estimating experimental error.

28.6.3.6 Delineating the field layout

An appropriate field layout matches the experimental design to the field gradients existing in the nursery. For RCBs, plots within a block should be as similar as possible; this necessitates close attention to water and soil gradients. It is often effective to block on nursery beds because plots within a bed are similar. Sometimes, however, more similarity can be achieved by blocking across beds, combining plots in several adjacent beds into the same block. It is frequently necessary to skip over certain local areas in the nursery that are dissimilar to other plots being included in a block. For example, relatively narrow, low-lying areas might be dissimilar to any of the other plots within a block and, if so, should be excluded from the experiment; these serve as buffers and are ignored for the purposes of experimentation.

Often, different blocks are put in different nursery fields. This has the advantage of broadening the range of validity to the entire nursery and makes plots within a block (field) more similar, thereby reducing experimental error. Attention should also be paid to possible carry-over effects resulting from previous nursery treatments. For example, suppose one part of the nursery had been hydromulched and another part mulched with sawdust. Because these two treatments could have lasting carry-over effects, plots within a block should come from areas receiving only one of the prior treatments.

28.6.3.7 Outlining the analysis

At this point, the biometrician should outline the form of the analysis, usually by delineating the sources of variation and degrees of freedom in the analysis of variance. Are the underlying hypotheses and probable precisions associated with the F tests in line with experimental objectives? What are the biological and practical implications of finding either significant or nonsignificant results for each test? If the chances are high that hypotheses will not be tested precisely enough, an alternate design is warranted. If the new design requires more effort or is not feasible, perhaps the experiment should be delayed or cancelled.

28.6.3.8 Documenting the plan

The experimental plan is often documented in the form of a research proposal or study plan by outlining the problem objectives and proposed experimental design as already described (see 28.6.1 to 28.6.3). This allows peer review, aids analysis, and documents the experiment in case of personnel turnover. Methods of writing study plans vary widely, depending upon the level of formality required.

In addition to the written description, the experimental design itself and the field layout are best documented by an analysis of variance table and a schematic diagram of the field plot arrangement. The table describes the form of initial analysis and succinctly states hypotheses under investigation. The schematic diagram, essentially a map, shows the layout of the treatments as they have been randomized and assigned to nursery plots; often, the positions of measurement plots within experimental plots are shown, as are any local areas omitted from the experiment. These diagrams can be simple or detailed; cryptic schematics are shown in Figures 3 to 6. The schematic (1) reinforces the written description of the experimental design by explicitly depicting the assignments of treatments to plots from which the analysis of variance can easily be constructed, (2) is useful during the experiment for applying treatments and collecting data, and (3) allows plot means to be charted as they occur in the field, often revealing spurious local trends.

The importance of documenting the experimental plan cannot be overemphasized; yet, too often, the effort involved hinders executing the research. Each researcher must find the most suitable compromise. Handwritten notes on the objectives and the experimental plan, including a list of factors and variables to be investigated, and a schematic map of the field plot layout are a *minimum* .

28.6.4 Conducting the experiment

28.6.4.1 Employing proper technique

Employing proper technique means conducting the experiment in a manner maximizing both precision and accuracy. High accuracy is achieved by using properly calibrated machinery and experienced, observant workers with proven good judgment and by closely following the experimental plan. Precision is increased by uniform application of treatments, meticulous measurement technique, and, in general, care and common sense.

28.6.4.2 Using the experimental design

Return, for a moment, to the concepts of randomization and blocking (see 28.5). When possible, seedlings should be treated, lifted, and measured according to the randomization scheme documented in the schematic diagram. As an extreme breech of this, consider the bias potentially introduced by first lifting and measuring all replicates of treatment 1. then treatment 2, and so on. As the experiment progresses, lifting conditions may change and measurement techniques become more refined. These effects can be randomized over treatments, thereby minimizing bias among treatment comparisons, by adhering to the original randomization scheme in all phases of the experiment.

To maximize the benefits from blocking, treat and measure all plots within a block before moving to other blocks. If seedlings in an entire experiment cannot be sown, lifted, or measured on the same day, do different blocks on different days. Then, any day-to-day differences in conditions tend to average out, influencing all plots (treatments) within a block similarly. When possible, one team of observers should lift and measure all seedlings in plots within a block. If one worker lifts

carelessly and measures trees in inches instead of centimeters, experimental accuracy and precision will be affected; however, any errors introduced will *not* bias treatment comparisons if that worker lifts and measures all treatments within a block because all treatments will have received the same poor technique.

28.6.5 Interpreting experimental results

After editing the data to eliminate data-collection errors, descriptive statistics (point estimates) must be calculated from the data set and inferences made about the population of interest (see 28.5.1). The analysis itself is beyond the scope of this chapter. But if the experiment is properly designed and conducted, the descriptive statistics should be unbiased, precise estimates of population parameters, and the level of uncertainty associated with those estimates and with tests of hypotheses should be low. Nevertheless, interpreting the practical implications of these statistical tests and inferences raises some problems and is nearly always the province of the researcher, not the biometrician. Three problems of interpretation are considered in this section.

28.6.5.1 Statistical significance

The nursery researcher must always interpret the practical significance of experimental results in light of their statistical significance. However, three different situations may arise in which the researcher relies on deductive reasoning based on knowledge and personal experience to question or even ignore the statistical inferences obtained from data analysis.

First, statistically significant differences among treatments may be too small to be practically important. This implies that the experiment was more sensitive (powerful) than required for that particular hypothesis or treatment comparison. For instance, if doubling the amount of N fertilization results in a statistically significant (say, at the 99% confidence level) increase in average 2+0 seedling height of 0.5 ± 0.2 cm, then the experiment was extremely sensitive. Though confident that this increase was not due to chance, the researcher may still decide that the small increase does not warrant the extra cost of the additional fertilizer.

Second-the reverse of the first case—a treatment comparison or hypothesis may not be statistically significant; yet the magnitude of the differences involved may be biologically or practically significant. Suppose that fall N fertilization of 2+0 seedlings results in a 20% increase in root-growth capacity and a 40% increase in early-winter frost hardiness, compared to the controls. If neither of these differences approaches statistical significance, one of two alternatives exists. Either the variable nature of the traits has resulted in large differences occurring by chance or a Type 2 error (β) is being made. Recall that Type 2 errors result when the experiment is not powerful enough to declare differences even though they, in fact, exist. When differences of practical importance are not statistically significant (say, at $\alpha = 0.05$), the experimenter can calculate the magnitude of differences required to approach statistical significance. If values of frost hardiness must differ by 100%, the researcher would question the power of the experiment and perhaps plan a better one.

Third, a statistically significant result may contradict biological principles or past experiences. In this situation, (1) results may be spurious (on the average, 1 out of 20 tests at $\alpha = 0.05$ will be incorrectly declared significant), (2) treatments may have been mislabeled, or (3) the biological reasoning may be flawed. The experimenter must be open to all eventualities and reexamine both the planning and conducting of the experiment and the biological theory underlying it to see where the fault lies. Sometimes, the statistically significant difference has a low range of validity—for example, when two treatments declared statistically different were tested in only one part of the nursery for a single growing season. This experimental design may lead to spurious results, especially if plot location or growing season were atypical. Such tests of significance require scrutiny.

28.6.5.2 Correlation versus causation

Experiments are most often conducted to establish cause-effect relationships of practical significance; that is, the presence of a certain level of a factor under investigation causes an identifiable response in a measured variable. The experiment is set up to determine these cause-effect relationships by specifically controlling the factor levels. However, correlations among variables not being controlled may also be found during experiments, and while useful, these must be interpreted with extreme caution.

For example, an experiment testing different levels of N fertilization may indicate that fertilized seedlings are significantly taller than controls. If soil P levels are also measured (but not controlled) on each plot, there may be a strong, statistically significant correlation indicating that high soil P levels are associated with faster growth. However, it is *not* correct to conclude that higher P levels cause faster growth. P may not be a limiting element for growth at all but may simply be indicative of (a proxy variable for) the level of organic matter on a plot. If more organic matter causes faster growth and produces more soil P, then the correlation between P and growth will be high even though no causality exists between the two variables. The pitfall of inappropriately assigning causality to such correlations cannot be overstressed.

28.6.5.3 Interactions

Statistically significant interactions among factors are common in the biological sciences. When more than two factors are involved, the practical interpretation of the interaction may be difficult; however, proper interpretation of two-factor interactions is essential to drawing correct conclusions from nursery experiments. A range in magnitude of these interactions can exist, and three different types are considered here.

For a two-factor experiment investigating the effects of three levels of fall N fertilization and two levels of irrigation on early-winter frost hardiness of 2+0 Douglas-fir seedlings, three hypothetical outcomes (Fig. 9) are considered. Suppose that low water levels have the consistent effect of "shutting seedlings down," causing them to enter dormancy early, and therefore increasing early-winter frost hardiness. If N aids this metabolic transition independent of water level, then no interaction between N and water exists (Fig. 9a). Now suppose that increasing N levels increases early-winter frost hardiness regardless of water level, but the rate of increase is faster when water levels are high. That is, N is more effective in the presence of high water levels (perhaps because the additional water is needed for N to better aid the physiological transition). This type of interaction—in which two factors affect each other, but trends within each factor are similar when plotted over a second factor—is called a scale effect (Fig. 9b). Note that the practical interpretation in both cases a and b is very similar: high N and low water levels result in superior early-winter frost hardiness.

But when one factor acts differently in the presence than in the absence of a second factor, levels of the first factor change their ranking, depending upon the level of the second factor. For example, high N levels may be more effective in the presence of high water levels, but low N levels may be more effective in the presence of low water levels (Fig. 9c). Perhaps too much N "burns" the seedlings and retards frost-hardiness development if seedlings are not well watered. This type of interaction, called rank change, makes it impossible to describe the effects of N without considering the particular water level applied and greatly alters the conclusions drawn from the experiment.

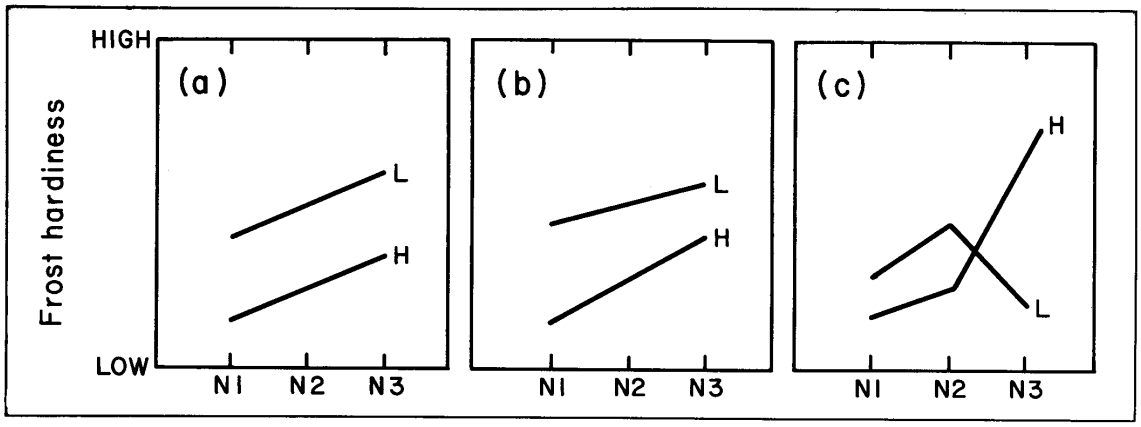

Figure 9. Hypothetical effects of two factors, irrigation level (H = high, L = low) and fall nitrogen fertilization (NI, N2, N3), on early-winter frost hardiness of 2+0 Douglas-fir for (a) no two-factor interaction, (b) scale-effect Interaction, and (c) rank-change Interaction.

28.7 Conclusions and Recommendations

The main goal of nursery research is to develop new techniques that produce high-quality seedlings in a cost-effective manner. After the initial steps of identifying a problem area and setting experimental objectives, the nursery researcher plans an experiment, by choosing treatments to test and variables to measure, determining plot sizes, selecting an appropriate experimental design, determining the number of replications, and delineating the field layout (assigning treatments to plots). The experiment is then conducted in a manner to maximize precision and accuracy of experimental results. Finally, the results are analyzed and interpreted in light of the researcher's intuition and personal experiences, and recommendations are made to implement the conclusions.

Applying a very few statistical concepts (mainly randomization, replication, and blocking) in a common-sense manner can aid researchers at each step of the nursery research process. While implementing the design and interpreting the results, the researcher must always balance statistical with biological and practical considerations to achieve an experiment that meets its objectives with an appropriate expenditure of effort.

Acknowledgments

I sincerely thank the following people for their help with and technical reviews of this manuscript: Mary Duryea, Oregon State University; Jim Fischer, U.S.D.A. Forest Service; Henry Laik, International Paper Company; Tom Landis, U.S.D.A. Forest Service; Paul Morgan, State of Oregon; Susan Stafford, Oregon State University; and Barbara Thompson, International Paper Company.

References

1. Anderson, S., A. Auquier, W. Hauck, D. Oakes, W. Vandaele, and H. Weisberg. 1980. Statistical methods for comparative studies. John Wiley and Sons, New York. 289 p.
2. Anderson, V. L., and R. A. McLean. 1974. Design of experiments: a realistic approach. Marcel Dekker, Inc., New York. 418 p.
3. Andrew, C. O., and P. E. Hildebrand. 1977. Planning and conducting applied research. MSS Information, New York. 116 p.
4. Bancroft, T. A. 1964. Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. Biometrics 20:427-442.
5. Cochran, W. G. 1976. Early development of techniques in comparative experimentation. Pages 3-25 *in* On the history of statistics and probability (D. B. Owen, ed.). Marcel Dekker, Inc., New York.
6. Cochran, W. G., and G. M. Cox. 1957. Experimental designs. John Wiley and Sons, New York. 61 1 p.
7. Cox, D. R. 1958. Planning of experiments. John Wiley and Sons, New York. 308 p.
8. Freese, F. 1967. Elementary statistical methods for foresters. U.S.D.A. Forest Serv.. Washington, D.C. Agric. Handb. 317. 87 p.
9. Gomez, K. A. 1972. Techniques for field experiments with rice. International Rice Res. Institute, Los Banos, Philippines. 46 p.
10. Hamburg, M. 1974. Basic statistics. Harcourt Brace Jovanovich, Inc., New York. 451 p.
11. Kempthorne, O. 1973. Design and analysis of experiments. Robert E. Krieger, New York. 631 p.
12. LeClerg, E. L., W. H. Leonard, and A. G. Clark. 1962. Field plot technique. Burgess Publishing Co., Minneapolis, Minnesota. 371 p.
13. Little, T. M., and F. J. Hills. 1972. Statistical methods in agricultural research. Agric. Extension, Univ. of California, Berkeley. 242 p.
14. Mason, E. G. 1981. Improvement of nursery and establishment techniques. Pages 186-198 *in* Forest nursery and establishment practice in New Zealand (C. G. R. Chavasse, ed.). New Zealand Forest Serv., Rotorua. FRI Symp. 22.
15. Neter, J., and W. Wasserman. 1974. Applied linear statistical models. Richard D. Irwin, Inc., Homewood, Illinois. 842 p.
16. Sandquist, R. E., P. W. Owston, and S. E. McDonald. 1981. How to test herbicides at forest tree nurseries. U.S.D.A. Forest Serv., Pacific NW Forest and Range Exp. Sta., Portland, Oregon. Gen. Tech. Rep. PNW-127. 24 p.
17. Snedecor, G. W., and W. G. Cochran. 1967. Statistical methods. Iowa State Univ. Press, Ames. 593 p.
18. Steel, R. G. D., and J. H. Torrie. 1960. Principles and procedures of statistics. McGraw-Hill Book Co., New York. 481 p.
19. Westmeyer, P. 1981. A guide for use in planning and conducting research projects. C. C. Thomas, Springfield, Illinois. 116 p.