J. Range Manage.
45:285-290, May 1992

Viewpoint: Replication, randomization, and statistics in range research

DAVID B. WESTER

Author is assistant professor, Department of Range and Wildlife Management, Texas Tech University, Lubbock 79409.

Abstract

Appropriate application of significance tests in statistical analyses requires an explicit statement of hypothesis; a clear definition of the population(s) about which inferences are to be made; and a model, a sampling strategy, an analysis, and an interpretation that are consistent with these considerations. In particular, experimental design and analyses must recognize appropriate replication and random selection of experimental units from target population(s). This paper discusses some aspects of these issues in range science research. Textbook examples and examples from range science applications are discussed in parallel in an attempt to clarify issues of randomization and replication in statistical applications.

Key Words: experimental design, pseudoreplication, sampling, inference

Hurlbert's (1984) monograph on pseudoreplication in ecological studies has encouraged scientists in many disciplines to examine field research problems from the critical perspective of experimental design as related to replication. Notwithstanding the importance and timeliness of Hurlbert's paper, it has also generated considerable confusion for many. Numerous studies that can be appropriately analyzed statistically have been accused of pseudoreplication because of an unclear understanding of the population(s) about which inferences are intended, what constitutes an experimental unit of the target population(s), and how these considerations apply to randomization and replication.

My purpose is to briefly discuss some aspects of the "problem of pseudoreplication" as it is commonly encountered in range science research. I begin with a brief discussion of some concepts related to analysis of variance, analysis of regression, replication, and randomization. Textbook examples and examples from range science applications are then discussed in parallel in an attempt to clarify issues of randomization and replication in statistical applications. My intention is not to be critical of any given author or research project; therefore, the range-related studies described below are real but fictitious names are used.

Discussion

It is helpful to begin by distinguishing between "data analysis and interpretation" and "statistics". These terms are not synonymous labels for identical endeavors. Tukey and Wilk (1966, p. 695) stated that:

> "The basic general tenet of data analysis is simply stated: to seek through a body of data for interesting relationships and information and to exhibit the results in such a way as to make them recognizable to the data analyst and recordable for posterity.

Statistics, on the other hand, is based on

> formal theories...[that 'legitimize'] variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions (in which a bare minimum of adjustable constants deny almost all flexibility) and [restore] the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with 'known' probabilities of error " (Tukey and Wilk 1966, p. 695).

These 2 prominent statisticians remarked that "While many of the influences of statistical theory on data analysis have been helpful, some have not" (Tukey and Wilk 1966, p. 695), and suggested that "Data analysis can gain much from formal statistics, but only if the connection is kept adequately loose" (Tukey and Wilk, 1966, p. 696).

To be too narrowly focused on the formal theoretical application of mathematical statistics is as undesirable as to be totally unconcerned with the proper application of statistics in the broader endeavor of experimentation and data analysis (e.g., Box 1978, p. 265–266, 270–271). A balance can perhaps be best achieved by clearly specifying the objectives and hypotheses motivating a study: by this process, a scientist makes explicit the population(s) and types of inference involved. This first step in research lies at the foundation of Eberhardt and Thomas' (1991) classification and discussion of field experiments. The following discussion is restricted to applications of formal experimental design considerations.

Once a population is defined, a statistical model is chosen to represent the behavior of the dependent variable as a function of explanatory variables. Assuming an appropriate model is selected, then certain conclusions about hypotheses can be made and extended to the population if research methodology properly incorporates replication and randomization assumed by the model and analysis. The degree to which the research has successfully addressed its objectives and hypotheses must be judged in light of the definition of, and correspondence between, the *population*, the *model*, the *sampling methodology*, the *analysis*, and the *conclusions* that are drawn.

Analysis of Variance and Analysis of Regression Models

Although "analysis of variance" and "analysis of regression" are often regarded as different analyses, they are actually different aspects of the same basic analysis of a linear model. The fundamental difference between analysis of variance and regression lies in the nature of the explanatory portion of the model (the independent variables, or the "expectation function"), and hence, appropriate application of these techniques depends upon the objectives of the research, the nature of the data collected, and the hypotheses to be tested. In "analysis of regression," the independent variables are generally continuous variables, and each experimental unit may have a different value for the independent variable. The analysis is usually couched in terms of fitting a line (or plane) through a scatter of points. Hypotheses commonly relate to the slope of the line (or plane). Prediction of the dependent variable for specified values of the explanatory variable(s) is also an important application.

In "analysis of variance," the independent variables are "dummy" or "class" variables whose designation and meaning depend upon the underlying group structure of the data. Observations on the dependent variable that share the same values of the independent

This is contribution T-9-541-000, College of Agricultural Sciences, Texas Tech University. The author wishes to thank C. Bonham, C. Britton, W. Laycock, A. Matches, H. Mayland, G. McPherson, R. Murray, A. Rasmussen, C. Scifres, and H. Wright for comments on an earlier draft of this paper.
Manuscript accepted 4 September 1991.

variable(s) are assumed to be random samples from the group designated by the class variables. Hypotheses generally relate to population means for groups under study. Equality of population means is the most commonly tested hypothesis.

Replication and Randomization

Examples discussed in this paper are relatively simple in that they involve only 1 treatment factor (or in the regression example, 1 independent variable). With these examples it is easier to focus on 4 components of experimental design which, when clearly defined, not only characterize the design but also guard against pseudoreplication. These components are the: (1) population(s) to which inference is extended, (2) treatment(s) under study, (3) experimental units that are treated, and (4) randomization rule used in the assignment of treatments to experimental units. Concepts of randomization and replication apply to single factor completely randomized designs as well as to more complex designs involving restriction in randomization (e.g., randomized block designs and latin square designs), factorial and split plot treatment arrangements, and repeated measures analyses.

Randomization and replication are both necessary for the appropriate application of significance testing in experimental designs. As Eberhardt and Thomas (1991, p. 55) stated, "Confirmation that two experimental outcomes are indeed different depends on randomization and replication to provide a measure of variability in units treated alike." Replication provides an estimate of experimental error: treatment differences are judged in light of the inherent variability among experimental units treated alike. Randomization plays a different but complementary role (Cox 1980, p. 313):

> randomization provides the physical basis for the view that the experimental outcome of a given study is simply one of a set of many possible outcomes. The uniqueness of the outcome, its significance, is judged against a reference set of all possible outcomes under an assumption about treatment effects, such as such effects are negligible. For the logic of this view to prevail, all outcomes must be equally likely, and this is achieved only by randomization.

If statistical concepts related to, for example, alpha level and the power of a test to detect real differences are to be interpreted explicitly, then these interpretations will be valid only insofar as the assumptions underlying the concepts are satisfied. Thus, if a scientist wishes to claim that "treatment differences are significant at the 5% level," then that scientist should also be willing to "pay the freight" for that statement: The scientist should design and conduct the study in accordance with principles upon which the inferences are based.

These concepts are illustrated in the following 4 "case studies." Study 1 involves analysis of regression for prediction purposes. Discussion is restricted to linear models for the sake of simplicity and familiarity. It should be noted that assumptions associated with significance tests in nonlinear models are usually identical to those associated with significance tests in linear models (Bates and Watts 1988). Study 2 illustrates analysis of variance for treatment mean comparison. A situation suitable for contingency table analysis is described in Study 3. The discussion of each of these studies is presented in 2 parts: part (a) is a "classical" description of the analysis from a widely used text, and part (b) is an analagous application from a range science setting. Study 4 discusses an analysis similar to Study 2 but with correlated errors.

Comparison of "Classical" and "Range-Related" Examples

Study 1, A

The relationship between serum cholestrol (the dependent variable Y) and age (the independent variable X) in women is studied (Snedecor and Cochran 1980, p. 385–388). Random samples of 56 women from Iowa and 130 women from Nebraska are selected. Linear regression is used to describe the relationship between Y and X for women in each state. The experimental unit is the individual woman. It is assumed the women selected in each state are representative of the populations to which inferences are directed. The model $Y_i = \beta_o + \beta_1 X_i + e_i$ is assumed for women in each state. The portion of the model on the right hand side of this equation that does not include the errors, e_i, is referred to as the "expectation function" (Bates and Watts 1988). The formal statistical hypothesis, H_o: $\beta_1 = 0$, addresses whether a linear relationship exists between Y and X.

The fundamental statistical assumptions for significance testing in this model generally offered by texts are that the errors, e_i, are independently, identically, and normally distributed with homogeneous variances (e.g., Graybill 1976, Theorem 6.3.1, p. 189–191). If errors have heterogeneous variances, variance-stabilizing transformations may be helpful. Although most theoretical studies of this model have assumed normality of errors, the F test is relatively robust to violations of this assumption (Pearson 1931, Lunney 1970). Probably the most important assumption is independence of errors. Random and mutually exclusive sampling often allows one to analyze data as though the assumption of independence is satisfied (Ostle 1963, p. 249–250). Correlated error structures (see Study 4) may be analyzed by modifying the calculation of the F statistic (Graybill 1976, p. 207–212, Smith and Lewis 1980, Pavur and Lewis 1983, Scariano et al. 1984, Scariano and Davenport 1984); failure to apply such modifications may have a large impact on type I error rates (Smith and Lewis 1980, Scariano et al. 1984). Three other assumptions of this model are: (1) the expectation function is correct, (2) the dependent variable is in fact equal to the expectation function plus the error, and (3) the errors are independent of the expectation function. Bates and Watts (1988) provide an excellent discussion of these assumptions [also see Steinberg and Hunter 1984)]. Application of simple linear regression analysis is appropriate when X is subject to measurement error if the primary objective of the research is prediction of Y (Sokal and Rohlf 1981, p. 549). This analysis is also appropriate when X and Y have a bivariate normal distribution (Steel and Torrie 1980, p. 246).

In addition to within-state analyses in Snedecor and Cochran's example, further hypotheses may address a comparison of the relationship between Y and X between the 2 states. For example, are the models for the 2 states identical? If not, then is the slope of the regression line the same for the 2 states? Is the Y-intercept the same for the 2 states? These questions may be appropriately addressed through statistical analyses such as comparisons of simple (Graybill 1976, Theorems 8.6.1–8.6.3, p. 288–291) and general (Graybill 1976, Theorem 8.6.4, p. 291–293) linear models. These tests are presented in the context of analysis of covariance in applied texts (e.g., Snedecor and Cochran 1980).

Several features of this study are noteworthy. First, only one value of Y need be observed at a given value of X in each sample of women. For example, in the sample from Iowa only 1 woman may have X=33. The relationship between Y and X for each sample can be estimated with only one observation on Y given X under the familiar assumption that each realized Y value is a random sample from a normally distributed population of Y's (at a corresponding value of X), and that each population of Y's has the same variance. It is important to note that statistical literature describes this situation as an application of linear regression that lacks replication (e.g., Scariano et al. 1984). When several independent observations on Y given X are available, then it is possible not only to explicitly test these assumptions, but also to examine the "lack of fit" of the model to the data (e.g., Montgomery and Peck 1982, p. 75–79).

Second, there is only 1 sample of (randomly selected) women

from each state. It is obvious that these data are adequate and appropriate to estimate regression equations within each state. However, it is also possible with these data to compare the relationship (between *Y* and X) between states with respect to slope, intercept, or linear combinations of slope and intercept (Graybill 1976, Theorem 8.6.3, p. 289–291). As Graybill (1976, p. 283) states: "For instance, an investigator is studying 3 different experimental situations and assumes a linear model for each. He wants to determine if these 3 linear models are dentical, or he may want to determine if some of the parameters of the models are the same from model to model." Graybill (1976) develops the theory to test such hypotheses. That is, 1 regression line has been estimated for each experimental situation, and it is possible to compare these lines. These tests and their interpretations are well defined. It is not unreasonable to suggest that the research may be "stronger" from a biological perspective if several samples of women from each state were used because a larger sample would be available to estimate population parameters. Nevertheless, only 1 sample of women from each state is required from a strict statistical viewpoint, and it is assumed that the collection of experimental units comprising each sample is both representative of its target population and large enough to obtain reasonable estimates of population parameters. Insuring representativeness calls into play appropriate sampling techniques and research methodology. Whether the sample is large enough depends on the precision in estimation desired by the researcher. An important factor is the ratio of the sample size (n) to the number of parameters being estimated (p); common recommendations suggest that the ratio n:p be from 30:1 to 400:1 (e.g., Kerlinger and Pedhazur 1973).

Study 1, B

Scientist White studied the relationship between redberry juniper (*Juniperus pinchotii* Sudw.) canopy cover (X) and herbaceous production *(Y)* on 2 upland sites characterized by redberry juniper-mixed grass vegetation in western Texas. Two populations, or experimental situations (*sensu* Graybill 1976, p. 283), were of interest: an ungrazed (relict) area and a heavily grazed area. Study locations were an isolated butte that was inaccessible to domestic livestock and a nearby grazed area. These 2 study sites were within 10 km of each other and had similar landscape position, underlying substrate, and soils. The primary objective was to describe the relationship between juniper canopy cover and herbaceous production on each site. Inference was intended to apply to these 2 sites. Other study areas may have different relationships between canopy cover and herbaceous production for a variety of reasons (e.g., edaphic conditions, past management history, etc.).

An individual juniper stand ("stand" was defined by the researcher) was considered an experimental unit. Fifty juniper stands were randomly selected on each site; in each stand, juniper canopy cover and herbaceous production were estimated with belt transects and clipped quadrats, respectively. Because grazing history of each site was well known, experimental units were treated similarly within each site. The analogy to Study 1, A is clear: whereas Snedecor and Cochran described a relationship between serum cholestrol and age in randomly selected women from Iowa and Nebraska, White studied the relationship between herbaceous production and juniper canopy cover in randomly selected stands on a relict site and a grazed site. Under usual statistical assumptions, both studies are amenable to statistical analyses. In particular, regression lines can be estimated within states (or sites) as well as compared between states (or sites).

A common misunderstanding of White's research relates to what has been incorrectly referred to as pseudoreplication. In particular, 2 specific criticisms are frequently directed at White's research. First, it has been claimed that because only 1 relict area and 1 grazed area were studied, it is not possible to compare the regression lines estimated for each site. Second, it is sometimes suggested that the comparison of the relationship between juniper canopy cover and herbaceous production between "grazing treatments" is completely confounded with the 2 sites studied. These criticisms are readily answered with the reminder that these 2 sites *are the* populations of interest: the comparison between these 2 sites is not intended to apply to grazed and relict sites in general. It is clear from Graybill (1976, p. 283–302) that such a comparison is statistically valid: 2 (or more) independent regressions *can* be compared.

These criticisms concern the definition of population and the representativeness of the samples of their respective populations. The populations in question in Study 1 are the 2 states (Iowa and Nebraska), or the 2 areas (relict and grazed). The experimental unit is the woman (in Snedecor and Cochran) or the stand (in White). If the collection of experimental units (as a random sample from a well-defined population) is representative of its population, then the study is amenable to appropriate statistical analyses. If, on the other hand, it is not reasonable to assume that the collection of experimental units is representative of its target population, then the research is flawed to the extent that inferences from the sample to the population are unwarranted because the former is not representative of the latter.

If the 2 regressions are shown to differ, for example with respect to slope, then the attribution of this difference to a grazing effect is an interpretational issue that is best dealt with from an ecological perspective. The F test may be used to show a difference between regressions, each representing an "experimental situation" in Graybill's (1976, p. 283) sense. The ecological interpretation of this difference involves information related to, for example, elevation, precipitation, and edaphic characteristics as well as differences in grazing history. It may be that the difference between regressions is due to some factor other than grazing. If this is so, then attributing the difference to a grazing effect is faulty because of confounding from an ecological viewpoint. To explicitly and exclusively incorporate grazing history as the "treatment" factor responsible for any differences that may be detected changes both the scope of the research and the experimental design requirements necessary to address the hypotheses in question. In particular, this new objective can be satisfied by conducting the research on more than 1 relict site (each representing a random sample from the population of relict sites) and more than 1 grazed site (each representing a sample from the population of grazed sites).

Study 2,A

Gomez and Gomez (1984:13–17) describe a study of the effect of chemical control of brown planthoppers and stemborers on rice yield. Treatments (6 chemical and a control) are randomly assigned to 4 replications (plots) each; that is, the experimental unit is the individual plot. Gomez and Gomez (1984:2–4) provide straightforward discussion of the need both for replication and randomization, and state unambiguously that "to obtain a measure of experimental error [the difference among experimental units treated alike] replication is needed." The issue of subsampling was discussed in a later chapter (Gomez and Gomez 1984, p. 241–255); for our purpose, subsampling may be incorporated into the present example by assuming that 5 randomly located quadrats in each plot are harvested to estimate rice yield. Therefore, an appropriate statistical model for this design is $Y_{ijk} = \mu + \tau_i + e_{(i)j} + \epsilon_{(ij)k}$. In this model, τ_i represents the i'th treatment effect, $e_{(i)j}$ is the experimental error associated with the j'th replication of the i'th treatment, and $\epsilon_{(ij)k}$ is the sampling error associated with the k'th sample in the j'th replication of the i'th treatment. The variance of $e_{(i)j}$ is denoted o_e^2 and the variance of $\epsilon_{(ij)k}$ is denoted o_s^2.

Based on this model that includes not only treatment effects, but

Table 1. Analysis of variance table for Studies 2,A and 2,B.

Source of variation	Degrees of freedom	Study 2,A	Study 2,B	Expected mean square
Treatment	$(t-1)$	6	2	$\sigma_s^2 + s\sigma_e^2 + sr\sum_{i=1}^{t}\tau_i^2/(t-1)$
Experimental error	$t(r-1)$	21	0	$\sigma_s^2 + s\sigma_e^2$
Sampling error	$tr(s-1)$	112	12	σ_s^2
Total	$trs-1$	139	14	

also distinguishes between experimental error (variation between experimental units treated alike) and sampling error (variation within experimental units), these data may be summarized in an analysis of variance table (Table 1). Experimental error can be estimated when r > 1. Based on expected mean squares, this estimate of experimental error is used to evaluate treatment effects (also see Steel and Torrie 1980, p. 155). In particular, the hypothesis H_o: $\mu_1 = \mu\ \mu_2 = \ldots = M_7$ can be tested, where μ_i is the mean rice yield in the i'th treatment. Assumptions underlying this F test include normality and independence of experimental errors as well as homogeneity of variances of experimental errors among treatments. Variation between experimental units relative to variation within experimental units may be evaluated to address issues of sampling efficiency (Cochran 1977). Pooling experimental error and sampling error is discussed by Paull (1950), Storm (1962) and Gill (1978).

Case Study 2, B

Scientist Black studied the effect of prescribed fire in redberry juniper-mixed grass vegetation in the Texas Rolling Plains. The objective of the research was to develop management recommendations for the vegetation type. Fire treatments were applied to pastures 800–1,200 ha in size; the study examined a 4-year old burn, an 8-year old burn, and an unburned control pasture. Each treatment (age of burn) was represented by 1 pasture. Habitat was evaluated in part by examining vegetation structure and composition. Shrub cover was estimated along 5 randomly located 100-m line transects in each pasture. Frequency of herbaceous species was recorded in 10, 0.5-m^2 quadrats randomly located along each transect line. Hypotheses of interest included whether mean shrub canopy cover and frequency of selected forb species differed between burning treatments.

In this study, the experimental unit (the unit of experimental material to which a treatment is applied) was an individual pasture. Three pastures were randomly assigned to burning treatments. Despite the fact that there were 5 sampling units (line transects) per pasture (and 50 quadrats per pasture for herbaceous frequency), there was only 1 replication (pasture) per treatment. Based on an analysis of variance (Table 1), there is no information to estimate experimental error because experimental error degrees of freedom are 0. The only estimate of error available is sampling error, and in view of the expected mean squares in Table 1, use of sampling error to evaluate treatment effects leads to a biased F test when $\sigma_e^2 > 0$ (also see Steel and Torrie, p. 155).

Black's study is a straightforward example of what Hurlbert (1984) called "simple pseudoreplication." As Hurlbert (1984, p. 201) noted, "multiple samples per experimental unit do not increase the degrees of freedom available for testing a treatment effect." It is often claimed that significant differences between unreplicated treatments may be attributable to the treatment effect if, prior to treatment, differences between plots were found to be statistically nonsignificant. Hurlbert (1984, p. 200–201) dispels any misunderstanding on this point with a detailed example showing how simple pseudoreplication increases the probability of detecting spurious treatment effects. Hurlbert (1984, p. 200) states:

> The validity of using unreplicated treatments depends on the experimental units being *identical* at the time of manipulation and on their remaining *identical* after manipulation, except insofar as there is a treatment effect. The lack of significant differences prior to manipulation cannot be interpreted as evidence of such identicalness. This lack of significance is, in fact, only a consequence of a small number of samples taken from each unit.

Hurlbert (1984, p. 203) suggests "The question to be asked is not: 'Are experimental units sufficiently similar for one to be used per treatment?' Rather it is: 'Given the observed or expected variability among experimental units, how many should be assigned to each treatment?'"

It was stated earlier that Scientist Black's objective was to develop management recommendations for burning redberry juniper vegetation. That is, the population comprises the entire vegetation type. The pastures used in Black's study represent that population. However, with only 1 pasture per burning treatment, there is no measure of variability between experimental units treated alike, and thus no information is available to extend inferences about treatment effects to other pastures (i.e., to the population).

It is possible in Black's study to compare mean canopy cover among his particular pastures with a redefinition of the populations to which inference is to be extended (and hence a redefinition of the model applied to the data as well as the hypothesis to be tested). One may regard each pasture as a population and each transect as a sample from that population. Variation from transect to transect then estimates variation inherent in the population. An F test may be used to compare mean canopy cover *among these 3 pastures*. This approach is being used more and more commonly in applied ecological research (e.g., Belsky 1986, Guthery 1987, Schulz and Guthery 1987, Thurow et al. 1988, Baker and Guthery 1990, Dormaar et al. 1990).

Study 3, A

Steel and Torrie (1980, p. 500–501) discuss an example of the effect of enrichment of bacterial inoculum with different vitamins on mortality of mice. Treatments are inoculum cultured in broth with 4 different vitamins. The experimental unit is an individual mouse; 9–13 mice are assigned randomly to treatments. The response variable is mouse survival or death. Data may be arrange in a 2 × 4 contingency table, with "dead" and "alive" as row designations and the 4 treatments as column designations. The hypothesis that the proportion of live mice does not differ between treatments may be tested with a chi-square test or a likelihood ratio G test. It is important to emphasize that although these nonparametric tests are distribution-free tests, they are not assumption-free tests. In particular, assumptions underlying these tests are that: (1) the observations in each treatment are a random sample from the corresponding population, (2) the observations in the treatments are mutually independent, and (3) each observation may be classi-

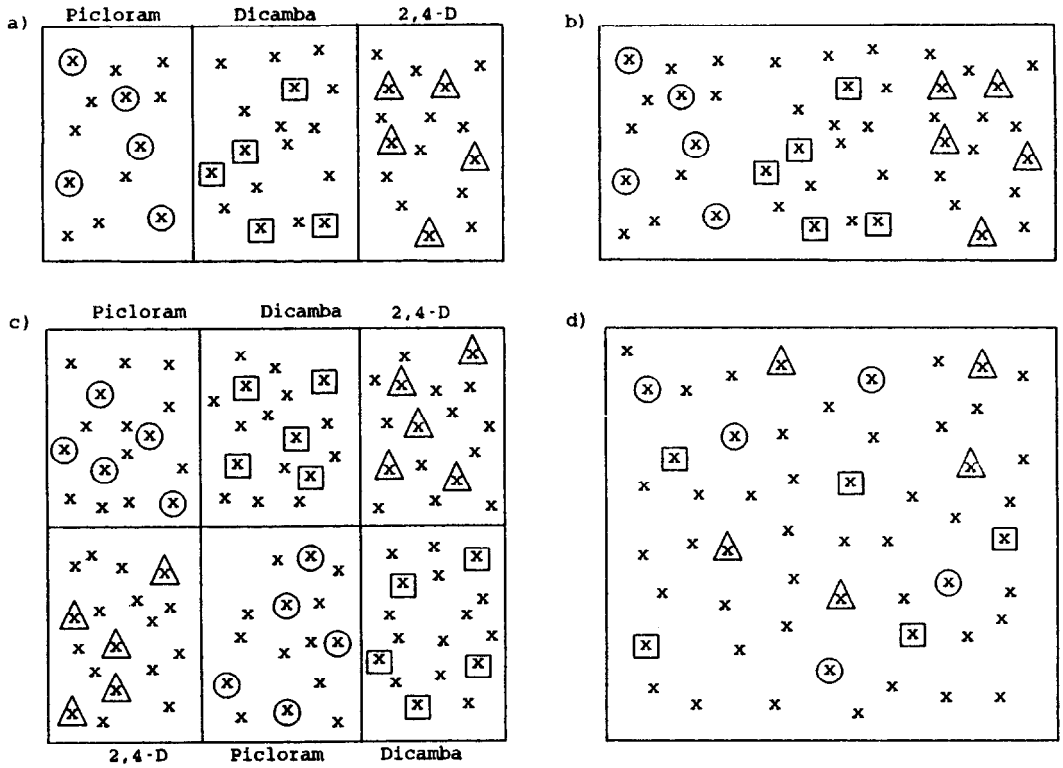

Fig. 1. Plot diagrams for Study 3, A. Trees are represented by x's; circled x's are treated with picloram; x's with squares are treated with dicamba; x's with triangles are treated with 2,4-D. (a) Each rectangle represents a 0.4-ha plot. Chemical are assigned to plots, with 1 plot per chemical. (b) Chemicals are (nonrandomly) assigned to trees rather than to plots. (c) Similar to part (a) except there are 2 plots randomly assigned to each chemical. (d) Similar to part (b) except that trees are randomly assigned to chemicals.

fied into the categories "dead" or alive" (Bishop et al. 1975, Conover 1979).

Study 3, B

Scientist Gray studied the effect of 3 herbicide treatments on mesquite (*Prosopis glandulosa* Torr.) mortality. The field design was as follows: 3 contiguous 0.4-ha plots were established and randomly assigned to a chemical treatment, with 1 plot per treatment. In each plot 5 trees were randomly selected and hand-sprayed with the chemical assigned to the plot.

The study can be described in 2 ways (Fig. 1). First, suppose the experimental unit is the 0.4-ha plot; this is, in fact, the unit to which the chemical treatment was assigned (Fig. 1a). Mortality would be expressed as the percentage of plants killed in each plot. Data would be summarized in contingency table with 2 rows (dead and alive) and 3 columns (corresponding to treatment). If the experimental unit is defined as the 0.4-ha plot, then this study lacks treatment replication: the approach of expressing mortality as percentages yields only 1 datum per treatment. If there had been at least 2, 0.4-ha plots per treatment (and assignment of chemical treatment to plots was random), then mortality among treatments could be tested for equality with an F test (after appropriate transformation) because each plot (experimental unit) yields an estimate of mortality. With only 1 plot per treatment, however, treatments are randomly assigned to experimental units (0.4-ha plots), but treatments are not replicated.

A second approach would be to consider the individual tree as the experimental unit; there would then be 5 experimental units per treatment (Fig. 1b). Data would be summarized in a contingency table with 2 rows (dead and alive) and 3 columns (corresponding to treatment). However, this approach also has problems. If the individual tree is considered the experimental unit, then there is replication but treatments are not randomly assigned to experimental units. Each experimental unit (tree) in the study area did not have an equal chance of receiving a treatment: even though individual trees were sprayed, treatment (chemical) assignment was to the (group of trees in the) 0.4-ha plot; hence treatments were not randomly assigned to experimental units. It is possible that there may be some systematic variation from plot to plot due to, for example, soil conditions, and differences in tree response from plot to plot may be attributable largely to one or more of these other factors and not to the chemical treatment. Snedecor and Cochran (1980, p. 127) provide a firm reminder: "Before claiming that the significant difference is caused by the variable under study, it is the investigator's responsibility to produce evidence that disturbing factors of this type could not have produced the difference. . .the device of randomization. . .makes it easier to ensure against misleading conclusions from disturbing influences."

This research could have been designed differently in 2 ways to render analyses appropriate. If the experimental unit is the 0.4-ha plot, then there should be at least 2 plots per treatment, and treatment assignment to plots should be random (Fig. 1c). Analysis of variance is appropriate to test the hypothesis that mortality does not differ among treatments; an arcsin transformation may be required because data are percentages. Alternatively, if the experimental unit is the individual tree, then several trees must be treated with each chemical, and again treatment assignment to experimental units must be random (Fig. 1d). A contingency table-based analysis with a chi-square or likelihood ratio G test is appropriate to test the hypothesis that the proportion of dead trees does not differ among treatments.

Study 4, A

Range research involving livestock often is subject to limitations and difficulties not typically encountered in researching dealing with plants. For example, consider a supplementation study using

steers on native rangeland. The treatment is level of protein supplementation. Suppose that 4 levels of supplementation are randomly assigned to animals, with 30 animals in each supplementation group. Due to feeding logistics, the 30 animals assigned to each treatment are randomly assigned to a fenced pasture. The 4 pastures used in the study are similar in all reasonable respects (e.g., similar management history, soils, forage type and availability, etc.).

In this study, the experimental unit for the supplementation treatment is the individual steer. With 4 treatments randomly assigned to 30 animals each, there is appropriate randomization of treatments to experimental units as well as appropriate replication. The model assumed for this study would be the linear model for a 1-way analysis of variance without a sampling error (see Study 2).

However, by keeping the 30 animals in each treatment in separate pastures, there is nonrandom handling of treatment groups. Two potential consequences of this are: (1) confounding of pasture effects with treatment effects, and (2) positively correlated errors within treatment groups due to a common environment (Gill 1978, p. 20). Whereas the first consequence may be reduced by selecting pastures that are as similar as possible, it is clear from Hurlbert's (1984, p. 201) statement regarding multiple samples per experimental unit that this is not an effective solution to the problem.

The second consequence (correlated errors) can have very serious implications in significance tests. One of the assumptions in the F test underlying the linear model for a completely randomized design is that the errors are uncorrelated. Failure to recognize and adjust for correlated errors changes the type I error rate of the F test. Smith and Lewis (1980) developed an adjustment for the F statistic to account for equicorrelated errors, and Scariano and Davenport (1984) extended the adjustment to cases with nonequicorrelated errors. Without these adjustments, actual significance levels differ from the nominal significance levels.

Conclusions

In many cases it is difficult to replicate or randomize appropriately, perhaps because of logistic constraints or financial limitations. Although these considerations often exercise profound influence on research, they do not excuse the rules of mathematical statistics. It may well be difficult or impossible to incorporate replication and randomization in the assignment of treatments to experimental units; however, this is not justification for interpreting and presenting results from statistical tests *as if* the study had been designed with appropriate replication and randomization. Eberhardt and Thomas (1991) present a detailed discussion of data analysis and interpretation of field studies which vary in the degree of control over which an investigator has in experimental design.

It is worthwhile to recall Hurlbert (1984, p. 188): "the quality of an investigation depends on more than good experimental design, so good experimental design is no guarantee of the value of study." The fact that appropriate replication and randomization are not incorporated into a study does not mean, in and of itself, that the study lacks useful information. The importance of proper experimental design as a key component in the process of "strong inference" (Platt 1964) is in no way diminished by recognizing that knowledge can be accumulated through an amalgamation of observational data and experimentation. The former information is not secondary to the latter, and scientific advancement may be more rapid and efficient if observational and experimental endeavors are used in a complementary way (e.g., Schoener 1983, Hawkins 1986).

Literature Cited

Baker, D.L., and F.S. Guthery. 1990. Effects of continuous grazing on habitat and density of ground-foraging birds in south Texas. J. Range Manage. 43:2–5.

Bates, D.M., and D.G. Watts. 1988. Nonlinear regression analysis, Its applications. John Wiley and Sons, N.Y.

Belsky, A.J. 1986. Revegetation of artificial disturbances in grasslands of the Serengeti National Parks, Tanzania. I. Colonization of grazed and ungrazed plots. J. Ecol. 74:419–437.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. Discrete multivariate analysis, Theory and practice. MIT Press, Cambridge.

Box, J.F. 1978. R.A. Fisher, The life of a scientist. John Wiley and Sons, N.Y.

Cochran, W.G. 1977. Sampling techniques. 3rd ed., John Wiley and Sons, N.Y.

Conover, W.J. 1979. Practical nonparametric statistics. John Wiley and Sons, N.Y.

Cox, D.R. 1980. Design and analysis of nutritional and physiological experimentation. J. Dairy Sci. 63:313–321.

Dormaar, J.F., S. Smoliak, and W.D. Willms. 1990. Distribution of nitrogen fractions in grazed and ungrazed fescue grassland Ah horizons. J. Range Manage. 43:6–9.

Eberhardt, L.L., and J.M. Thomas. 1991. Designing environmental field studies. Ecol. Monogr. 61:53–73.

Gill, J.L. 1978. Design and analysis of experiments in the animal and medical sciences. Iowa State Univ. Press, Ames.

Gomez, K.A., and A.A. Gomez. 1984. Statistical procedures for agricultural research, 2nd ed., John Wiley and Sons, N.Y.

Guthery, F.S. 1987. Guidelines for preparing and reviewing manuscripts based on field experiments with unreplicated treatments. Wildl. Soc. Bull. 15:306.

Graybill, F.A. 1976. Theory and application of the linear model. Duxberry Press, New Scituate.

Hawkins, C.P. 1986. Pseudo-understanding of pseudoreplication: a cautionary note. Ecol. Soc. Amer. Bull. 67:184–185.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54:187–211.

Kerlinger, F.N., and E.J. Pedhazur. 1973. Multiple regression in behavioral research. Holt, Rinehart and Winston, Inc., N.Y.

Lunney, G.H. 1970. Using analysis of variance with a dichotomous dependent variable: an empirical study. J. Educ. Measurements 7:263–269.

Montgomery, D.C., and E.A. Peck. 1982. Introduction to linear regression analysis. John Wiley and Sons, N.Y.

Ostle, B. 1963. Statistics in research, 2nd ed. Iowa State Univ. Press, Ames.

Paull, A.E. 1950. On a preliminary test for pooling mean squares in the analysis of variance. Ann. Math. Stat. 21:539–556.

Pavur, R.J., and T.O. Lewis. 1983. Unbiased F tests for factorial experiments for correlated data. Commun. Statist.-Theor. Meth. 12:819–840.

Pearson, E.S. 1931. The analysis of variance in cases of non-normal variation. Biometrika 23:114–133.

Platt, J.R. 1964. Strong inference. Science 146:347–353.

Scariano, S.M., and J.M. Davenport. 1984. Corrected F tests in the general linear model. Commun. Statist.-Theor. Meth. 13:3155–3172.

Scariano, S.M., J.W. Neill, and J.M. Davenport. 1984. Testing regression function adequacy with correlation and without replication. Commun. Statist.-Theor. Meth. 13:1227–1237.

Schoener, T.W. 1983. Field experiments on interspecific competition. Amer. Natur. 122:240–285.

Schulz, P.A., and F.S. Guthery 1987. Effects of short duration grazing on wild turkey home ranges. Wild. Soc. Bull. 15:239–241.

Smith, J.H., and T.O. Lewis. 1980. Determining the effects of intraclass correlation on factorial experiments. Commun. Statis.-Theor. Meth. 9:1353–1364.

Snedecor, G.W., and W.G. Cochran. 1980. Statistical methods, 7th ed., Iowa State Univ. Press, Ames.

Sokal, R.R., and F.J. Rohlf. 1981. Biometry, 2nd ed., W.H. Freeman and Co., San Francisco.

Steel, R.G.D., and J.H. Torrie. 1980. Principles and procedures of statistics, 2nd ed., McGraw-Hill Book Co., N.Y.

Steinberg, D.M., and W.G. Hunter. 1984. Experimental design: review and comment. Technometrics 26:71–104.

Storm, L.E. 1962. Nested analysis of variance: review of methods. Metrika 5:158–183.

Thurow, T.L., W.H. Blackburn, and C.A. Taylor, Jr. 1988. Some vegetation responses to selected livestock grazing strategies, Edwards Plateau, Texas. J. Range Manage. 41:108–114.

Tukey, J.W., and M.B. Wilk. 1966. Data analysis and statistics: An expository review. AFIPS Conf. Proc., Fall Joint Comp. Conf. 29:695–709.