Ecology, 78(5), 1997, pp. 1312–1320

BLOCKING FACTORS AND HYPOTHESIS TESTS IN ECOLOGY: IS YOUR STATISTICS TEXT WRONG?

Jonathan A. Newman,[1] Joy Bergelson,[2] and Alan Grafen[3]

[1]*Department of Zoology, Southern Illinois University Carbondale, Illinois 62901-6501 USA*
[2]*Department of Ecology & Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637 USA*
[3]*Department of Plant Science, University of Oxford, South Parks Roads, Oxford, OX1 3RB United Kingdom*

Abstract. We demonstrate that statistics textbooks differ in their prescription for the analysis of experiments that involve blocking factors. The differences in analysis may lead to differences in conclusions regarding the significance of experimental treatment effects. We outline the two approaches, discuss why they are different, and suggest when each approach may be applicable. We point out that simply following one's textbook may not be the best course of action for any particular situation.

Key words: blocking factors; error terms; hypothesis tests; mixed-model ANOVA; random effects; statistical inference.

Introduction

The modern ecologist must have a good working knowledge of statistics. But where do ecologists learn statistics? Typically they take statistics classes in departments such as psychology, mathematics, or even agricultural science. We are, for the most part, practitioners of statistics, not statisticians. We accept what we are taught and try to apply what we have learned correctly. But what if what we have learned is wrong? If one statistics book told us to conduct a test in one way, and another text told us to conduct it in a different way, would we know which is correct?

In this paper, we point out that there are two accepted approaches to the ''proper'' analysis of experiments incorporating blocking factors. Textbooks on statistics generally include only one of these tests and, in fact, rarely mention the existence of the other. But one cannot substitute perfectly for the other. The two tests make very different assumptions and allow very different inferences to be drawn. We run the risk of misinterpreting our data if we blindly follow the test outlined in our favorite statistics text. In this paper, we attempt to explain and contrast the two methodologies for the analysis of blocking factors in ecological experiments. We wish to emphasize that our intent is not to advocate either approach, but rather to provide the background necessary for an informed decision regarding which test to use. Indeed, the point of this paper is that there are two different, accepted approaches, each of which is well supported by volumes of statistical literature.

What are Blocking Factors?

The experimental unit is the unit to which a treatment is applied. It may be a plot of land, a population, or an individual plant or animal. Most biologists would expect these things to vary intrinsically in any variable we cared to measure. For instance, some plots of land will be more productive than others and some plants will grow larger than others, even in the absence of any differences in treatment. The aim of blocking is to group experimental units so that they are more similar within blocks than between blocks. By accounting for these intrinsic differences among our experimental units, we hope to obtain a smaller experimental error and hence improve the precision with which we estimate the treatment effect.

Blocking factors differ from treatments in that they have not, or cannot, be applied randomly to the experimental units. Statistically, blocks are thought of as a source of nuisance variation; biologically, this need not be the case. Indeed, ecologists often have a great deal of interest in the importance of the blocking factor. In many cases authors do not even call these factors ''blocks.'' Nevertheless, these factors are not randomly assigned to the experimental units. Regardless of our scientific interest in these factors, they are intrinsic sources of variation, that when grouped as blocks serve to remove this variation before we test our hypotheses about those treatments that have been applied experimentally.

Before delving into the important differences behind the treatment of blocking factors in ecology, we first review the rationale behind ANOVA and the differences between fixed and random effects. For herein lie the roots of the problem.

Manuscript received 2 November 1995; revised 4 November 1996; accepted 3 December 1996. For reprints of the Special Feature, see footnote 1, p. 1299.

The Rationale of ANOVA

An F ratio is a ratio of two independent random variables that each have a χ^2 distribution. We can use the F ratio as a test statistic for testing the null hypothesis that the variances from two populations are equal. We do this by drawing two independent samples, one from each population. For example, suppose we have $X = x_1, x_2, \cdots, x_m$ and $Y = y_1, y_2, \cdots, y_n$. We would, for each sample, calculate the sum of the squared deviations from the sample mean, which we usually call SS_x and SS_y. We could then form the ratio:

$$\hat{F} = \frac{\text{SS}_x/(m - 1)}{\text{SS}_y/(n - 1)} = \frac{\text{MS}_x}{\text{MS}_y}. \quad (1)$$

While this looks like the *F ratios* that we are used to seeing in ANOVA, we still have not shown that $\hat{F}$ has an F distribution. To do this we note that

$$F^* = \frac{\text{SS}_x/[(m - 1)\sigma_x^2]}{\text{SS}_y/[(n - 1)\sigma_y^2]} \quad (2)$$

follows an F distribution. This is because the numerator and the denominator can be shown to have χ^2 distributions (see for example, DeGroot 1986:386–387). Now, note that $\hat{F} = (\sigma_x^2/\sigma_y^2)\, F^*$. If $\sigma_x^2 = \sigma_y^2$ (i.e., the null hypothesis is true), then $\hat{F} = F^*$ and $\hat{F}$ has an F distribution. This simple logic is used in all ANOVA analyses, where as in Eq. 1 the ratio is formed from two mean squares.

ANOVA statistical designs specifically isolate the variance due to a treatment from the variance due to everything else in the experiment. Roughly speaking, we get something like:

$$\frac{E[\text{MS}_x]}{E[\text{MS}_y]} = \begin{array}{l}(\text{variance due to the treatment} \\ \quad + \text{ variance due to other things}) \\ \div\ (\text{variance due to other things}),\end{array} \quad (3)$$

where $E[\text{MS}_i]$ is the expected mean square. As long as the ''other things'' that make up part of the expected mean square in the numerator are the same as those that make up the denominator, this ratio tests the null hypothesis that the variance due to the treatment is zero. If the ''other things'' are not the same, then any statistical difference inferred from the ratio of these two mean squares may not necessarily be due to the treatment effect. In order to test the null hypothesis that the variance due to any particular treatment is zero, the correct denominator must be chosen, namely the one that contains the same ''other things.'' So, under the null hypothesis, these two mean squares will be approximately equal and $\hat{F}$ will be ≈ 1.

Fixed vs. Random Effects

In order to choose the ''correct'' denominator for a test, we must calculate the expected mean squares; to do this it is necessary to determine whether our factors are fixed or random effects. The distinction between the two types of effects probably began with Eisenhart (1947) but has been fully developed by many statisticians (see Yates 1965, 1970 for further critical discussion of this development). Here we briefly define what is usually meant by these terms.

A fixed effect is something that is repeatable (i.e., if another scientist wished to repeat our experiment, the same levels of the factor could be used) and the levels used in the experiment must represent all of the levels in the universe about which we are attempting to draw statistical inferences. For example, if we choose a factor D with five levels for an experiment, and we treat this factor as a fixed effect, then any inferences that we draw regarding the effect of factor D are applicale only to these five levels. As an example of treatment D, consider a species of pseudoscorpion that has five instars in its life history. If we conducted an experiment in which we used all five levels of ''instar,'' then ''instar'' should be considered a fixed effect because all possible instars are represented and another experimenter could use these same levels. If, instead, we conducted an experiment on only three levels of ''instar,'' and if we treated ''instar'' as a fixed effect, then our inferences would be restricted to this subset of its life history.

A random effect is an effect where the levels of the factor, call it G, are thought to be a random sample from an essentially infinite set of possible levels. When the levels of a treatment cannot be exactly replicated by another experimenter, then that treatment should almost always be thought of as a random effect. However, even treatments that are repeatable may be considered random effects if they can be regarded as a random sample from some larger population. The population from which the levels form a random sample is the universe about which inferences can be drawn. In other words, our inference extends beyond the data to the population of treatment levels as a whole. As an example, imagine that we use several offspring from each set of parents in an experiment and that we regard these parents as a random sample from some larger population. Since it would be impossible for someone to replicate the families used in the experiment, and since we wish to draw an inference regarding the entire population of families (rather than these particular families), then ''family'' would be treated as a random effect.

Unfortunately, a typical situation in ecology often fits neither scheme with respect to the blocks. On the one hand, the blocks cannot be replicated. On the other hand, they are usually not a random sample from any population. This kind of ambiguity is rarely dealt with in statistics textbooks.

One further note regarding random effects, which is often overlooked in statistics textbooks, is that inferences about the random effect are not inferences about the mean of the response (as they are with fixed effects),

TABLE 1. Effects of tethering on blue crab predation rates. This table illustrates two different approaches to the analysis of a single experiment that has two experimental treatments (tethering and habitat type), each with two levels. The treatments are applied to experimental units that have been grouped together by size class to form blocks. There were five size classes (instars 1, 3, 5, 7, and 9). These size classes form five blocks with each of the four treatment combinations replicated 11 times per block. Model 1 shows that there is a weak effect of habitat type, no effect of tethering, and no interaction. Model 2 concludes that there are very strong effects of habitat type and tethering but not their interaction. (It should be noted that Pile et al. [1996] used neither of these two models. They considered all sources to be fixed effects and calculated the sources seen in Model 1, but used the residual MS for all hypothesis tests.)

Source	df	SS	MS	F	
Model 1					
Blocks (B)	4	11.473	2.868		
Habitat (H)	1	0.751	0.751	10.01	$P = 0.034$
Tether (T)	1	0.360	0.360	3.53	$P = 0.134$
$H \times B$	4	0.301	0.075		
$T \times B$	4	0.409	0.102		
$H \times T$	1	0.004	0.004	0.12	$P = 0.739$
$H \times T \times B$	4	0.137	0.034		
Residual	180	8.458	0.047		
Total	199	46.53			
Model 2					
Blocks (B)	4	11.473	2.868	2.34	
Habitat (H)	1	0.751	0.751	5.63	$P = 0.000$
Tether (T)	1	0.360	0.360	5.22	$P = 0.007$
$H \times T$	4	0.004	0.004	0.69	$P = 0.765$
Residual	192	9.306	0.048		
Total	199	46.53			

they are inferences about the variance-covariance structure. For readers interested in pursuing this issue, Searle et al. (1992:9–12) have a nice discussion of this point.

WHAT ARE THE TWO APPROACHES?

The following example from Pile et al. (1996) will serve to illustrate some of these points. Among other things, Pile et al. were interested in testing experimentally whether the use of tethers on blue crabs provides an unbiased estimate of predation in newly settled crabs. These authors were also interested in whether predation rates differed between vegetated and unvegetated habitats. Because predation rates will vary among crabs of different sizes, even in the absence of treatment effects, Pile et al. grouped the crabs used in the experiment into five size classes. Depending on the assumptions that the authors make, two possible analyses are shown in Table 1. Both models consider blocks to be random effects and tethering and habitat to be fixed effects. The first model is prescribed, for example, by Edwards (1985:262) and the second is prescribed, for example by Mead (1988:37). We want to stress that we could easily cite dozens of textbooks that support either model; Table 2 provides a sample of how some commonly used textbooks treat the problem. These two models differ in their handling of the block by treatment interactions, for reasons that we will elaborate below. For the moment, it is sufficient to note that we would come to fundamentally different conclusions regarding the importance of tethering and the strength of the habitat effect.

As a means of motivating discussion, let us consider some more general examples of blocking in ecological experiments. These examples are useful in demonstrating the enormous variation in the types of blocking factors used by ecologists. Suppose that we are interested in the effects of water stress and nutrient stress on the growth of the plant ragwort and that we perform an experiment incorporating three levels of water stress and three levels of nutrient stress (i.e., nine treatment combinations). The following are four experiments that we might conduct to estimate the effects of the treatments.

Experiment A.—We seed a 18-ha piece of land with ragwort seeds. We divide the 18-ha piece into four equal-sized (4.5 ha) sections of land, which we further divide into nine 0.5-ha subsections. In each section, one of the nine 0.5-ha subsections are randomly assigned to each of nine water–nutrient combinations. After one year, we harvest all of the ragwort from each subsection and use the total dry mass from each plot (subsection) as our dependent variable. Thus, we have four sections of land (blocks), three levels of water stress, three levels of nutrient stress and one plot (subsection) of land per section–water–nutrient combination.

Experiment B.—We collect one seed from each of 36 ragwort mothers, and plant them in each of 36 planting pots. The seeds are divided into four groups based on the mass of the mother plant (high, medium-high, medium-low, and low). In each group, the nine seeds are randomly assigned to one of the nine water by nutrient combinations. After one year, each plant is harvested and the dry mass measured. Thus, we have four groups of plants (blocks), three levels of water stress, three levels of nutrient stress with one plant per group–water–nutrient combination.

Experiment C.—We obtain nine seeds from each of

TABLE 2. Initial belief about the block by treatment interaction. If your analysis is guided by any of these textbooks, then you would follow these approaches.

Interaction	No interaction
Bennett and Franklin 1954	Bowerman and O'Connell 1990
Blackwell et al. 1992	Lentner and Bishop 1993
Edwards 1985	Mead 1988
Kirk 1982	Mendenhall and Beaver 1991
Lentner 1993	Mendenhall and Sinich 1988
Potvin 1993	Neter et al. 1990
Villars 1951	Seber 1997
Winer 1970	Snedecor and Cochran 1989
Zar 1984	Sokol and Rohlf 1982
Zolman 1993	Steel and Torrie 1980

TABLE 3. Analyses of a hypothetical example. This table illustrates two different approaches to the analysis of a single experiment that has two experimental treatments (water stress and nutrient stress), each with three levels. The treatments are applied to experimental units that have been grouped together to form blocks. There are four blocks with each of the nine treatment combinations replicated once per block. Model 1 shows that neither of the main effects, nor their interaction is significant. Model 2 concludes that there are significant effects of water stress and nutrient stress but not their interaction. The last column shows the expected mean square for each term in each analysis. These are discussed in the *Existence of the interaction term* section.

Source	df	SS	MS	F		E[MS]
Model 1						
Blocks (B)	3	5.9	1.97			$\sigma^2 + wn\sigma^2_B$
Water stress (W)	2	9.45	4.73	4.43	NS	$\sigma^2 + n\sigma^2_{W\times B} + bn\theta^2_W$
Nutrient stress (N)	2	8.75	4.38	4.77	NS	$\sigma^2 + w\sigma^2_{N\times B} + bw\theta^2_N$
$W \times B$	6	6.4	1.07			$\sigma^2 + n\sigma^2_{W\times B}$
$N \times B$	6	5.5	0.92			$\sigma^2 + w\sigma^2_{N\times B}$
$W \times N$	4	2.3	0.58	0.84	NS	$\sigma^2 + \sigma^2_{W\times N\times B} + b\theta^2_{W\times N}$
$W \times N \times B$	12	8.23	0.69			$\sigma^2 + \sigma^2_{W\times N\times B}$
Residual	0	0	0			
Total	35	46.53				
Model 2						
Blocks (B)	3	5.9	1.97	2.34	NS	$\sigma^2 + wn\sigma^2_B$
Water stress (W)	2	9.45	4.73	5.63	$P < 0.01$	$\sigma^2 + bn\theta^2_W$
Nutrient stress (N)	2	8.75	4.38	5.22	$P < 0.05$	$\sigma^2 + bw\theta^2_N$
$W \times N$	4	2.3	0.58	0.69	NS	$\sigma^2 + b\theta^2_{W\times N}$
Residual	24	20.13	0.84			σ^2
Total	35	46.53				

Note: Lowercase letters denote the number of levels in treatments denoted by the corresponding uppercase letter (e.g., there are w levels in treatment W). The σ^2 symbols with subscripts denote the variance due to the treatments listed in the subscript, when at least one of those treatments is a random effect; σ^2 with no subscript denotes the error variance. The θ^2 symbols denote the variance due to the treatments listed in the subscript, when those treatments are all fixed effects.

four genetically different ragwort mothers. The seeds were mailed to us from a colleague, who cannot remember where he originally collected them. One seed from each mother is randomly assigned to each of the nine water-nutrient combinations. After one year, each plant is harvested and the dry mass measured. Thus, we have four mothers (blocks), three levels of water stress, and three levels of nutrient stress, with one plant per mother–water–nutrient combination.

Experiment D.—This experiment is identical to experiment C, except that the seeds were collected from a random sample of mothers growing in Albany county in upstate New York. Again, we have four mothers (blocks), three levels of water stress, and three levels of nutrient stress, with one plant per mother–water–nutrient combination.

The analyses of these experiments might follow either of the models shown in Table 3. The mechanics of each of these analyses are the same as those shown in Table 1. The analyses shown in Table 3 clearly come to different conclusions regarding the experimental treatments. Why are they different and which is correct?

THREE TYPES OF BLOCKS

Each experiment in our example uses a blocking factor. In experiment A, the 4.5-ha sections of land are the four blocks. In experiment B, the groups of seeds based on the mother's dry mass are the four blocks. In experiments C and D, the groups of seeds from each mother comprise the four blocks. We propose that these experiments exemplify three different types of blocks. In experiment A, the block size and the experimental unit are an arbitrary division of the experimental material; they could just as readily have been 9 ha and 1 ha instead of 4.5 ha and 0.5 ha. In experiment B, the block size was an arbitrary division, but there is a natural experimental unit size. We could have chosen to use only two groups of mother dry masses but individual plant is the most sensible experimental unit. In experiments C and D, there is a natural size to both the blocks and the experimental units (they are mother and offspring) and we have no choice in defining the sizes of either of these divisions.

WHAT IS THE DIFFERENCE BETWEEN THE MODELS?

The most apparent difference between the two analyses is the inclusion of the block by treatment interactions in the first model (see Tables 1 and 3). Including the interactions leads to two additional, important differences. The first difference is in the selection of the error term (denominator of the F ratio) for testing the experimental treatments. The second difference is in the amount of replication that is implied in the selection of the error term. Specifically, in the ragwort growth experiment (Table 3), under Model 1, the main effect of water stress is tested by $F_{2,6} = \text{MS}_W/\text{MS}_{W\times B}$, and the main effect of nutrient stress is tested by $F_{2,6} = \text{MS}_N/\text{MS}_{N\times B}$. This differs from Model 2 where the main effects of water stress and nutrient stress are both tested with the $\text{MS}_{\text{residual}}$ as the error term on 24 df. Note that the $\text{MS}_{\text{residual}}$ used in Model 2 is really just

$$\frac{\text{SS}_{W\times B} + \text{SS}_{N\times B} + \text{SS}_{W\times N\times B}}{6 + 6 + 12}.$$

That is, the $\text{MS}_{\text{residual}}$ is really just a ''pooling'' of all of the block by treatment interaction terms. This difference is even more extreme in the blue crab experiment (Table 1) where the effect of tethering is tested with 4 df in the denominator under Model 1 and 192 df under Model 2.

Note that some statisticians prefer a third model, that is very similar to Model 1, but differs in some of the technical assumptions. This model is not usually found in applied statistics textbooks, however it is implemented in some statistical software packages (for example SAS's PROC GLM). The differences between Model 1 and this third model are fairly technical and beyond the scope of this paper. The interested reader may wish to consult an advanced statistics text like Hocking (1985) for details of this test.

Existence of the Interaction Term

Choosing the appropriate denominator depends on the designation of factors as both fixed or random effects and the researcher's assumption regarding the additivity of the treatments. We must be clear on this subject; we cannot ever know, with certainty, whether the interaction exists. As unsatisfying as it might be, all that we can do is adopt one of two a priori attitudes toward the existence of the interaction: it is likely to exist, or it is unlikely to exist.

We return to our ragwort growth examples to consider the consequences of both of these a priori attitudes toward the interaction. For the discussion that follows, consider water stress and nutrient stress to be fixed effects. That is, we will draw inferences only about the levels of water stress and nutrient stress used in the experiment, and those levels may be repeated by other experimenters. Throughout the discussion that follows, we will be concerned with the existence of the treatment by block interactions.

Suppose that we consider it possible or likely that there is an interaction.—Then, we would include the interaction in the model (e.g., Model 1, Table 3). Following Edwards (1985), the expected mean squares are given in Table 3. The notation used in Table 3 is fairly standard and can be found in most advanced statistics textbooks. Briefly, σ^2 represents the variance in the random error associated with each experimental unit or observation. The lowercase italic letters represent the number of levels in the treatment denoted by the same capital italic letter (e.g., there are w levels in treatment W). θ^2_W is given by

$$\theta^2_W = \frac{\sum_{i=1}^{w} (\mu_i - \mu)^2}{w - 1},$$

where $\mu_i - \mu$ is the deviation from the mean caused by level i of treatment W. We note that when an effect is considered to be fixed, then

$$\left(\sum_{i=1}^{w} \mu_i - \mu\right) = 0. \qquad (4)$$

by definition. The same is true for nutrient stress, N, which is also considered to be fixed. Since the blocks are considered to be random, the restriction noted in Eq. 4 does not apply. We denote the variance attributed to blocks by σ^2_B rather than θ^2_B to show this difference.

It is easy to see that the proper F ratios for testing the main effects of water stress and nutrient stress are now formed using the corresponding block by treatment mean square as the error term. Similarly, Table 3 shows that if block by treatment interactions are assumed to be possible, then there are no appropriate error terms for testing the effects of the blocks or any of the block by treatment interactions themselves. This stems from the fact that the residual term is confounded with the three-way interaction (i.e., we have run out of degrees of freedom) in this particular example (but not, for example, in Table 1).

Suppose that we consider an interaction unlikely.—If we believe that the interaction is unlikely, then it is not included in the analysis (i.e., it is assumed to be zero). We then get the expected mean squares shown in the last column of Table 3. We see, as in Eq. 3, that the proper error term for forming an F ratio to test any effect in the model is the residual mean square.

Justifying The Difference

As we have demonstrated, either approach leads to a ''valid'' F test in the sense that the denominator isolates the component of variance that is of interest. Are they both correct? In the sections that follow, we further examine the concept of blocks and outline the rationale underlying their treatment by each of the two models. Both arguments are convincing, but very different.

Interpretation of Model 1

The rationale underlying Model 1 begins with the distinction between fixed and random effects in ANOVA and with the assumption that an interaction between treatment and block is possible or even likely. The basic beliefs are well illustrated by Edwards (1985):

> *If conclusions based on the outcome of the experiment are restricted to exactly the same blocks as those used in the experiment, the use of* $MS_{\text{W(C)}}$ [re-

> sidual mean square] *as the error mean-square for all tests of significance is justified. All too often, however, in summarizing the results of the research, the experimenter discusses the blocks as if they were random. The experimenter cannot have it both ways; the blocks are either fixed or they are random. If they are random, then MS_{BT}* [block by treatment interaction mean square], *not $MS_{W(C)}$, is the appropriate error mean-square for testing the treatment effects for significance.* —Edwards 1985:278

That is, Edwards assumes that an interaction is likely, includes it in the analysis, and advocates using the mean square for the block by treatment interaction as the error term for testing the effect of the treatment. This follows exactly from our treatment of random effects with the interaction, where the unit of replication for analyzing treatment effects was found to be the treatment by block interaction.

The logic of this approach is obvious when there is significant variation due to the interaction. However, when the interaction is zero, the interaction mean square (if it is included in the analysis) and the residual mean square estimate the same quantity. For this reason, many authors suggest that we test the interaction first and if there is no evidence of a significant interaction, then we should use the residual mean square to test the main effect. The residual estimate would be preferred because it is made with more degrees of freedom (i.e., it is based on a larger sample size). These same books further suggest that it is appropriate to pool the interaction with the residual sums of squares; after all, they represent estimates of the same thing. Thus, when there is no interaction because we have presented evidence that there is no interaction, then the appropriate error term is not necessarily the interaction mean square. Notice that this is exactly the logic of Model 2, except that there has been no test of the null hypothesis regarding the interaction.

How do we ‘‘know’’ that the interaction is zero? Followers of Model 1 adopt a cautious attitude by requiring only weak evidence (i.e., large α values) before accepting that there is an interaction. By requiring only weak evidence, the chance of a Type II error (failing to reject a false null hypothesis) is reduced. This requirement may appear to be rigorous, but actually provides us with a false sense of security. There are two reasons for this. First, the conditional nature of the test of the null hypothesis (i.e., the data determines which denominator will be used in the test) is not reflected in the *P* value associated with that test. Second, this procedure is tantamount to accepting the null hypothesis rather than failing to reject it. For these reasons, most textbooks point out that ‘‘some statisticians’’ feel that it is never appropriate to pool, or switch, denominators. Nevertheless, these same books invariably go on to describe methods for pooling. Hines (1995) has recently demonstrated that pooling interaction terms very rarely results in any appreciable increase in statistical power without undercutting the statistical validity of the analysis. We agree with Hines and feel that if the experimenter imagines that an interaction is possible then the interaction should be retained throughout the analysis (i.e., we feel it is inappropriate to pool).

Interpretation of Model 2

In this model, the block by treatment interaction is assumed to be nonexistent, and hence excluded from the analysis. As a consequence, the appropriate error term for the treatment effect is the residual mean square, regardless of whether we view blocks as fixed or random effects. This belief is advocated forcefully by Steel and Torrie (1980), who write:

> *Very often it is desirable to have blocks random when generalizations concerning treatment effects are to be made. . . . Residual mean-square is the appropriate error mean-square for testing block and treatment effects. . . .* —Steel and Torrie 1980

This approach is based on the attitude that the presence of a block by treatment interaction means that we can conclude that the treatment influences the response, but we can say little about how it does so. Some factors we have not controlled influence the effect of the treatment and until we discover what they are, we have little more to say. The presence of an interaction therefore threatens the whole rationale of the experiment, because advocates of Model 2 do not view their blocks as a random sample from any particular population.

The disappointing conclusion implied by the presence of the interaction may be moderated by two considerations:

1) Interaction means ‘‘nonadditive interaction.’’ If we can characterize a way in which blocks and treatments do combine simply, then we may rescue our experiment. For example, a log transformation would remove the interaction if the factors combined multiplicatively. However, until we can find some way of representing the effects of the treatment in a way that is independent of blocks, we do not understand the effect of the treatment.

2) Our prime interest may be in the direction of the treatment effects, and not in their magnitude. For example, even if the magnitude of the effect of water stress on seedling growth varies, it may well be of interest that it is always negative. In this case, we might test for a treatment effect in the analysis of variance that omits the interaction. This is reasonable when the absence of main effect would make an interaction highly implausible. For example, if water stress does not affect survival at all, then it cannot affect survival dif-

ferentially in different blocks. Essentially, we have added absence of an interaction to the null hypothesis of no treatment effect.

The Devil is in the Details

We saw in Tables 1 and 3 that these two models for the analysis of a single experiment might come to different conclusions, regarding the treatment effects. These two approaches differ on several levels.

1. They test different null hypotheses.—We are often sloppy about how we state our null hypothesis. We usually use phrases like ''there will be no effect of the treatment.'' Unfortunately, one source of confusion in the interpretation of these two approaches is that they actually test different null hypotheses. Model 1, by using the interaction mean square, tests the null hypothesis that the treatment may have an effect in each block, but averaged across the population from which this sample of blocks has been randomly selected, the effect is zero. On the other hand, Model 2, by using the residual mean square, tests the null hypothesis that there is no effect of the treatment in any block. The appropriateness of these different hypotheses should dictate which approach is used (see Yates 1965, 1970 for more on these differences).

We should note that although we have concentrated on hypothesis testing (as this is most relevant to ecologists), the same distinction can be made regarding estimation. That is, in constructing an estimate of a treatment effect, the standard error used depends upon which mean square is used as an estimate of the error.

2. The type of block has implications for the assumption of additivity.—Previously, we defined three types of blocks. Model 2 usually applies to situations where blocks and plots are arbitrarily defined. As in experiment A (above), the attitude that blocks are really just big plots and the logical criteria that plots are random and additive with treatments seem to justify this belief. That is, all statistical approaches assume that experimental units (plots) and treatments are additive. Otherwise, it would not be possible to define a treatment effect! (Note that we are taking about experimental units, not blocks.) Therefore, viewing blocks as ''large experimental units'' or ''large plots'' would lead one to the logical conclusion that blocks and treatments must be additive.

However, in experiments C and D it is perfectly reasonable to assume that there is experimental unit by treatment additivity but not block by treatment additivity. In these cases, a block by treatment interaction simply means that there is a gene by environment interaction. This difference strongly influences the initial assumptions that each model makes regarding the presence of the interaction. Followers of Model 2 initially assume additivity and require strong evidence (e.g., small α values) to the contrary before they are willing to give up this assumption. Followers of Model 1 concede beforehand that there is a strong possibility that there will be block by treatment interactions and hence require that evidence for an interaction be extremely weak (e.g., large α values) before they are willing to assume additivity. Parallel to these differences, advocates of Model 2 would regard the existence of a block interaction as a reason not to draw conclusions about the main effects. Advocates of Model 1 aim to draw conclusions about main effects in the presence of block by treatment interactions.

3. The type of block has implications in defining the population.—The third difference regarding blocks deals with the intended inference. The logic of applying a model that considers blocks to be random implies that we believe that there is a population of blocks over which we can make some generalization. Recall that the hypothesis being tested is that the average effect, in the population from which the sample of blocks was taken, is zero. In order for this ''average'' to be meaningful, there must be a population of blocks from which the blocks are a random sample.

Whenever the choice of blocks is arbitrary, it is fairly meaningless to draw inferences about the population from which blocks are a sample. For example, what would it mean to speak about the population of 0.5-ha plots at your field station, or a population of adult mass categories? For this reason, we are unlikely to wish to extrapolate beyond our designated blocks when interpreting the results of experiments A and B. Experiments C and D, on the other hand, both have blocks that form a sample from some natural population (i.e., ragwort mothers). In these examples, there is a population that defines parameters of interest. In experiment D, we know exactly what population was sampled and hence we may wish to make statistical inferences about its population parameters. In experiment C, however, the population defined by our sample is unknown. It is likely that they were not randomly sampled from any real population.

We must ask ourselves two questions: is there a population of interest? and, can we draw statistical inferences about that population? This is relatively easy to do in the case of experiments A and D, but experiments B and C are more problematic. Most textbooks do not distinguish types of blocks and therefore tend to give one of two answers: Model 1 or Model 2. Unfortunately, the choice is frequently unclear in practical applications.

4. Biological vs. statistical inference.—Much of the confusion regarding this problem revolves around how the results of an experiment are interpreted. That is, what has been shown statistically as opposed to biologically. This attitude is epitomized by Edwards' quote ''All too often, however, in summarizing the results of the research, the experimenter discusses the blocks as if they were random [after treating them as fixed]. The experimenter cannot have it both ways; the blocks are either fixed or they are random.'' In order to draw any inferences beyond the blocks used in the experiment, Edwards (1985) maintains that we must draw a statistical inference that results from treating blocks as random effects, including the interaction in the analysis and using the interaction mean square as the error term for testing the treatment effect.

The distinction between biological and statistical inference can be seen as follows. An experiment will often be performed at one experimental station. If nutrient stress was liable to have arbitrary effects in different locations, such experiments would not be worth publishing beyond the perimeter fence. However, such experiments are published more widely, and usefully so. The attitude adopted is that the experiment has demonstrated nutrient stress to have an effect in the station. Against a background of past success in generalizing from one area to another, and in the absence of any specific contrary information, it is reasonable to conclude tentatively that nutrient stress will have the same effect elsewhere. In other words, the statistical inference is restricted to where it can be clearly defended on logical grounds, and to admit that any further extension of the conclusions has no statistical justification.

Model 1 avoids completely any nonstatistical inference. If the blocks are something that have a natural unit and a natural population of those units, and we are prepared to make the assumption that our sample forms a random sample from that population, then on the basis of that assumption we can generalize statistically to the population from which the blocks were drawn. This ensures that every inference is drawn statistically. Nevertheless, the same logical uncertainties exist as before concerning the extension of experimental results, but now they have been incorporated into the assumptions of the statistical test. The degree to which the blocks form a random sample of a natural population of blocks is the degree to which our conclusions are justified statistically.

Every useful application of statistics makes unjustified assumptions about the similarity of the experimental material and the material about which conclusions are drawn. In Model 2, these unjustified assumptions are kept separate from the statistical conclusions so that the analysis itself provides safe but limited conclusions. In Model 1, these assumptions are drawn into the statistical argument, so that the analysis may provide widely applying conclusions but that may be applicable to an uncertain population.

Pseudoreplication Implications

In 1984, Stuart Hurlbert wrote an influential, and much needed, monograph on ''Pseudoreplication and the design of ecological field experiments'' (Hurlbert 1984). As a result of this paper, many ecologists now recognize that we should be cautious about defining our unit of replication, and that a failure to recognize when samples lack independence can lead to erroneous conclusions. Although not the focus of Hurlbert's paper, it is clear that pseudoreplication can be framed in terms of analysis of variance designs. Hurlbert remarks, almost in passing, that ''In ANOVA terminology, it [pseudoreplication] is the testing for treatment effects with an error term inappropriate to the hypothesis being considered.''

Pseudoreplication is relevant in the context of the present paper in that there are two different hypotheses of potential interest, and two potential error terms. If we are attempting to test the hypothesis that ''averaged over the population from which blocks are a random sample, the effect is zero despite the fact that treatments might have an effect in each block,'' then the appropriate error term is the block by treatment interaction, and the replication is the number of block by treatment combinations! Conversely, if we are testing the hypothesis that ''there is no effect of the treatment in any block,'' then the error term is the residual mean square and the replication comes from the number of experimental units that are nested within the block by treatment combinations. The effect of this difference can be seen clearly by referring to Table 3. In the first case, the test of the main effect of water stress is made with 6 df in the error term, while in the second case the test of the main effect of water stress is made with 24 df in the error term! The correct replication depends on the hypothesis that we are testing and on whether we have satisfied the assumptions of that test.

Conclusion

We have shown that choosing the correct denominator is particularly relevant when using blocking factors in the ANOVA and have presented two ''mainstream'' approaches to the treatment of blocking factors. Our presentation shows that the choice of models is not a black and white issue. The differences rest largely on the assumptions that we are willing to make, which in turn are influenced by the type of blocking factor that we are using. When blocks are arbitrarily defined, it is reasonable to assume that there is no block by treatment interaction. Thus, Model 2 would be adopted and our null hypothesis would be that there is no effect of the treatment in any block. However, when blocks represent natural units then the presence of block by treatement interactions is entirely plausible. In the absence of additional information to the contrary, we should retain the interaction in the analysis by adopting Model 1. In this instance, we are testing the null hypothesis that, when averaged across the population from which these blocks have been randomly selected, the treatment has no effect. Since many factors can influence our choice of statistical model, and since there is no one correct method for most specific experiments, the only reasonable action to adopt is to calculate all possible sources of variance and publish the ANOVA tables so that readers can decide for themselves what the experiment means, or whether adopting a different approach would lead to different conclusions.

Acknowledgments

This research was supported by a grant from the British Biotechnology and Biological Research Council to J. A. Newman (Joint Agriculture and Environment Programme, JAEP) and grants from the National Science Foundation (DEB 9350363) and the Packard Foundation to J. Bergelson. We gratefully acknowledge the support of Linacre College (J. A. Newman), and Worcester College (J. Bergelson), both of the University of Oxford. We thank P. Armitage, R. A. Bailey, C. Hitchcock, H. Jones, A. Kacelnik, P. Kareiva, J. Losos, R. Mead, T. Shaffer, and an anonymous reviewer for valuable discussion.

Literature Cited

Bennett, C. A., and N. L. Franklin. 1954. Statistical analysis in chemistry and chemistry industry. J. Wiley & Sons, New York, New York, USA.

Blackwell, T., C. Brown, and F. Mosteller. 1991. Which denominator? Fundamentals of exploratory analysis of variance. In C. Hoaglin, F. Mosteller, and J. W. Tukey, editors. J. Wiley & Sons, New York, New York, USA.

Bowerman, B. L., and R. T. O'Connell. 1990. Linear statistical models: an applied approach. Second edition. PWS-Kent, Boston, Massachusetts, USA.

DeGroot, M. H. 1986. Probability and statistics. Second edition. Addison-Wesley, Reading, Massachusetts, USA.

Edwards, A. L. 1985. Experimental design in psychological research. Fifth edition. Harper & Row, New York, New York, USA.

Eisenhart, C. 1947. The assumptions underlying the analysis of variance. Biometrics **3**:1–21.

Hines, W. G. S. 1996. Pragmatics of pooling in ANOVA tables. American Statistician **50**:127–134.

Hocking, R. R. 1985. The analysis of linear models. Brooks/Cole, Monterey, California, USA.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**: 187–211.

Kirk, R. E. 1982. Experimental design: procedures for the behavioral sciences. Second edition. Brooks/Cole, Monterey, California, USA.

Lentner, M., and T. Bishop. 1993. Experimental design and analysis. Second edition. VA, Valley Book Company, Blackburg, Virginia, USA.

Mead, R. 1988. The design of experiments: statistical principles for practical applications. Cambridge University Press, Cambridge, England.

Mendenhall, W., and R. J. Beaver. 1991. Introduction to probability and statistics. Eighth edition. PWS Kent, Boston, Massachusetts, USA.

Mendenhall, W., and T. Sinich. 1988. Statistics for engineering and the sciences. Third edition. Dellen, San Francisco, California, USA.

Neter, J., W., W. Wasserman, and M. H. Kutner. 1990. Applied linear statistical models. Third edition. Irwin, Homewood, Illinois, USA.

Pile, A. J., R. N. Lipcius, and J. van Montfrans. 1996. Density-dependent settler-recruit-juvenile relationships in blue crabs. Ecological Monographs **66**:277–300.

Potvin, C. 1993. ANOVA: experiments in controlled environments. In (S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Chapman and Hall, New York, New York, USA.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. Variance components. J. Wiley & Sons, New York, New York, USA.

Seber, G. A. F. 1977. Linear regression analysis. John Wiley & Sons, New York, New York, USA.

Snedecor, G. W., and W. G. Cochran. 1989. Statistical methods. Eighth edition. Iowa State University Press, Ames, Iowa, USA.

Sokal, R. R., and F. J. Rohlf. 1982. Biometry. Second edition. W. H. Freeman, New York, New York, USA.

Steel, R. G. D., and J. H. Torrie. 1980. Principles and procedures of statistics: a biometrical Approach. Second edition. McGraw-Hill, New York, New York, USA.

Villars, D. S. 1951. Statistical design and analysis of experiments for development research. W. M. C. Brown, Dubuque, Iowa, USA.

Winer, B. J. 1970. Statistical principles in experimental design. McGraw-Hill, New York, New York, USA.

Yates, F. 1965. A fresh look at the basic principles of the design and analysis of experiments. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: 777–790.

———. 1970. Experimental design: selected papers. Charles Griffin, London, England.

Zar, J. H. 1984. Biostatistical analysis. Second edition. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Zolman, J. 1993. Biostatistics: experimental design and inference. Oxford University Press, Oxford, England.