Anova for Unbalanced Data: An Overview

Ruth G. Shaw; Thomas Mitchell-Olds

Ecology, Vol. 74, No. 6. (Sep., 1993), pp. 1638-1645.

Stable URL:
http://links.jstor.org/sici?sici=0012-9658%28199309%2974%3A6%3C1638%3AAFUDAO%3E2.0.CO%3B2-G

Ecology is currently published by Ecological Society of America.

Ecology, 74(6), 1993, pp. 1638–1645
© 1993 by the Ecological Society of America

ANOVA FOR UNBALANCED DATA: AN OVERVIEW[1]

RUTH G. SHAW[2]
Department of Botany and Plant Sciences, University of California,
Riverside, California 92521 USA

THOMAS MITCHELL-OLDS
Division of Biological Sciences, University of Montana,
Missoula, Montana 59812 USA

Abstract. Ecological studies typically involve comparison of biological responses among a variety of environmental conditions. When the response variables have continuous distributions and the conditions are discrete, whether inherently or by design, then it is appropriate to analyze the data using analysis of variance (ANOVA). When data conform to a complete, balanced design (equal numbers of observations in each experimental treatment), it is straightforward to conduct an ANOVA, particularly with the aid of the numerous statistical computing packages that are available. Interpretation of an ANOVA of balanced data is also unambiguous. Unfortunately, for a variety of reasons, it is rare that a practicing ecologist embarks on an analysis of data that are completely balanced. Regardless of its cause, lack of balance necessitates care in the analysis and interpretation. In this paper, our aim is to provide an overview of the consequences of lack of balance and to give some guidelines to analyzing unbalanced data for models involving fixed effects. Our treatment is necessarily cursory and will not substitute for training available from a sequence of courses in mathematical statistics and linear models. It is intended to introduce the reader to the main issues and to the extensive statistical literature that deals with them.

ANOVA AND BOONS OF BALANCE

In this section we briefly review ANOVA, noting the advantages of a strictly balanced design. For single factor analyses, lack of balance does not present serious problems (Milliken and Johnson, 1984:127). We therefore discuss the two-way factorial design with the effects of both factors considered fixed, because it is among the simplest that reveals the main distinctions between balanced and unbalanced cases. The principles we present hold, in general, for more complicated models involving fixed effects. Consideration of mixed and random effects models, in which interest focuses on the variance of effects, rather than on estimates of the effects themselves, is appreciably more complex and, for this reason, is beyond the scope of this paper, but a treatment of this topic can be found in textbooks on the subject (Searle 1971, 1987, Milliken and Johnson 1984; see also Shaw 1987*a* for a consideration of analysis of quantitative–genetic data).

In the balanced two-way factorial design, there are two treatment factors (A and B), each having several different states or levels. All the possible combinations of the n_a levels of Factor A with the n_b levels of Factor B are generated, $n_a \times n_b = p$, and one treatment combination is applied to each of the $N = p \times r$ experimental units, where r is the number of observations per treatment combination, or cell. One or more measures (y) are taken on each experimental unit.

Given such a design, the *means model,*

$$y_{ijk} = \mu_{ij} + e_{ijk}, \tag{1}$$

where μ_{ij} is the mean of the ijth cell of the factorial design, e_{ijk} is the deviation of the kth observation in the ijth cell from the mean of that cell, $i = 1$ to n_a, $j = 1$ to n_b, and $k = 1$ to r, expresses the individual observations in terms of the cell means.

An alternative model, the *effects model,*

$$y_{ijk} = \mu + a_i + b_j + t_{ij} + e_{ijk}, \tag{2}$$

where μ is the grand mean, a_i (b_j) is the additive contribution of the ith (jth) level of Factor A (B) on the response, t_{ij} is the deviation of the mean of the ijth cell from the sum of the ith and jth marginal means, e_{ijk} is the deviation of the kth observation in the ijth cell from the mean of that cell, $i = 1$ to n_a, $j = 1$ to n_b, and $k = 1$ to r, describes the effects of the treatment states

[1] For reprints of this Special Feature, see footnote 1, p. 1615.

[2] Present address: Department of Ecology, Evolution and Behavior, University of Minnesota, St. Paul, Minnesota 55108 USA.

on the responses. Whereas the means model has p parameters, one for each cell mean, the effects model has $q = (1 + n_a + n_b + p)$ parameters. With balanced data, the effects model often corresponds to factors or concepts being put to experimental test. However, it may be necessary to rely on the means model for analysis of some unbalanced data sets. In either model, it is assumed, for the purposes of hypothesis testing, that the e_{ijk} are independent of one another and identically distributed in a Normal distribution having mean = 0, and some variance, σ^2.

Regardless of the choice of parameterization, either model can be expressed conveniently in matrix notation as:

$$\mathrm{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{3}$$

where y is the $N \times s$ matrix of s responses observed for each of N individuals, $\mathbf{X}$ is the $N \times r$ ($r = p$ or q, depending on the parameterization) "design matrix," $\mathbf{b}$ is an $r \times s$ matrix of the parameters of either of the two models, above, and $\mathbf{e}$ is an $N \times s$ matrix of residuals. In the univariate case, $s = 1$, and $\mathbf{y}$, $\mathbf{b}$, and $\mathbf{e}$ are column vectors. The design matrix contains known constants denoting the contributions of particular parameters to the expected value of an individual. We usually define the element $X_{kl} = 0$ otherwise. Note that the interpretation of the elements of $\mathbf{b}$ depends on the parameterization. The least squares estimate of the set of parameters, $\mathbf{b}$, is

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{4}$$

The superscripts, T and -1, indicate matrix transpose and generalized inverse (Searle 1971), respectively. This representation reveals that ANOVA can be viewed as a familiar problem of multiple regression analysis.

When the means model (Eq. 1) is used, the solution $\mathbf{b}$, consisting of the estimates of the cell means, is readily obtained. In contrast, the effects model (Eq. 2) is over-parameterized (that is, there are more parameters than can be estimated from the available information), so the $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ matrix is singular, and infinitely many solutions, $\mathbf{b}$, exist. This problem can be resolved by imposing restrictions on the parameters of Eq. 2 (Searle 1971, 1987, Milliken and Johnson 1984:Chapter 6). Differing restrictions produce distinct estimates of $\mathbf{b}$. Regardless of the choice of restrictions, however, identical estimates are obtained for the *estimable functions,* linear combinations of the model parameters that by definition do not depend on the restrictions set on the parameters. For example, in the effects model specified above for the two-way crossed design (Eq. 2), estimable functions for each of the p cell means are obtained by summing the elements of $\mathbf{b}$ that estimate each effect contributing to a given mean (i.e., $\mu + a_i + b_j + t_{ij}$). The important issues of estimability of the parameters in the effects model and estimable functions are considered in more detail by Milliken and Johnson (1984) and Searle (1987).

The investigator usually wishes to answer each of the following questions by testing the corresponding null hypothesis, H_0:

1) Does the effect of one factor on the response variable(s) depend on the level of the other factor? H_0: There is no interaction between Factor A and Factor B. This null hypothesis is expressed as $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0$ for all i, i', j, j', with $'$ indicating distinct states of a factor (means model), or as $(t_{ij} - t_{i.} + t_{.j} + t_{..}) = 0$ (effects model) (where . as a subscript indicates averaging over the levels of a given factor).

2) Do the levels of Factor A differ in their effects on the response variable(s)? H_0: There is no main effect of Factor A on the response. In the means model, this null hypothesis is given as $\mu_{1.} = \mu_{2.} = \ldots = \mu_{p.}$. The same hypothesis can be expressed in the effects model as all $(a_i + t_{i.})$ are equal.

3) Do the levels of Factor B differ in their effects on the response variable(s)? H_0: There is no main effect of Factor B on the response. This null hypothesis can be expressed in terms of the parameters as $\mu_{.1} = \mu_{.2} = \ldots = \mu_{.p}$ (means model) or $(a_i + t_{.j})$ are equal (effects model).

Note that all hypotheses can be expressed in terms of either the means model or the effects model. Available statistical packages are based on the effects model, but Milliken and Johnson (1984) demonstrate the utility of the means model, particularly for unbalanced designs. They also show how standard packages can be used to conduct ANOVA in terms of the means model.

The analysis of variance procedure is so named, because it breaks down ("analyzes") the variance (actually, the total sum of the squared deviations of the responses from their grand mean [i.e., $\mathrm{SS_T} = (\mathrm{N} - 1)$ times the variance]) into terms that quantify the magnitude of the overall variance in the response variable(s) attributable to the factors of the design, to their interaction, and to residual variability within cells of the design. There are several distinct methods for accomplishing the breakdown of the $\mathrm{SS_T}$ into the SS for the different factors. Computing formulae are available in many statistical texts (e.g., Milliken and Johnson 1984:Chapters 9 and 10) and need not be repeated here. In the balanced case, as defined above, the methods yield identical results, and interpretation is therefore straightforward. When the design is balanced, moreover, the sums of squares corresponding to each of the factors, to their interaction, and to the residual variance are independent of one another, and these sums of squares are distributed according to the noncentral χ^2 distribution with their respective degrees of freedom (df). In the cases we are considering, a model having only fixed effects, the "importance" of a given factor is *usually* judged by comparison of the variance attributable to that factor to the residual (within-cell) variability. Thus, tests of each null hypothesis are developed by constructing the ratio of the mean square (MS = SS/df) for the particular effect to the residual MS. (We urge caution here, however. With certain designs (e.g., nested, split-plot), tests of particular effects require a denominator other than the residual MS; see Milliken and Johnson 1984:Chapters 5 and 24–32). When the null hypothesis holds, then, given a balanced design, this ratio of MS is distributed exactly according to the F distribution. When the design is unbalanced, the distinct methods of partitioning SS_T do not produce the same results, the resulting SS associated with the two factors and their interaction are not necessarily independent of one another, and the ratio of MS is no longer exactly distributed according to the F distribution. Thus, loss of balance causes ambiguities that plague the processes of estimating the parameters, partitioning the SS, and testing the hypotheses of interest. Although this is certainly a discouraging situation, a careful analysis can often overcome these impediments and can provide a reasonably clear picture of the biology embodied in the data.

Types of Imbalance

We use the term "balance" to refer collectively to several distinct attributes of data structures. Balance can therefore be compromised in several different ways, which we describe in this section.

Given the balanced two-way factorial design described above, there are three ways that balance can be marred: (1) the numbers of observations for the different treatment combinations may be unequal, (2) some of the cells (treatment combinations) may be missing altogether, and, (3) in multivariate data, some of the experimental units may have been measured for only a subset of the response variables. We consider these in turn.

Unequal sample size

Probably the most common way in which data are unbalanced is by inequality of numbers of observations per cell. When the unit of observation is an individual organism, then mortality, emigration, or inability to relocate individuals for measurement can affect numbers representing a given treatment. Even when the unit of observation is a group of organisms, it is sometimes necessary to eliminate units due to accidents during application of the treatments or during measurement.

If properties of the treatments are likely to be causally related to the variation in sample size among cells, then analysis of only the available responses (e.g., of survivors), ignoring the missing observations, would reveal only part of the effect (or even obscure the effect) of the treatments applied. (See Little and Rubin 1987: 8–9, and Maxwell and Delaney 1989:273, for further discussion of this point.) One way of achieving a more complete picture of the overall response to the treatments is to use categorical methods to explicitly analyze the effects of the treatments on final numbers of individuals in each cell (e.g., Shaw 1986, 1987*b*). The necessity of separately analyzing the realized cell sizes and the responses measured on remaining individuals is unfortunate, because the two analyses cannot be regarded as independent, and no joint analysis is readily available. However, we are optimistic that current research in theoretical statistics (e.g., Little and Rubin 1987:Chapters 11 and 12) will eventually permit joint analysis of the pattern of missing data together with variables measured on the remaining experimental units. In the following, we assume that the treatments do not directly cause the variation in sample size or that the variation in sample size is analyzed separately.

Regardless of the cause of the variation in numbers of observations, its consequence is that there are more observations for some combinations of levels of the factors, and hence more information on the effect of these combinations, than for other combinations. That is, the levels of the factors, often called "independent variables," are *not* independent in the realized data. As a result, the estimates and tests of the effects of factors are also not generally independent. Thus, the lack of balance impairs the ability of the experiment to accomplish the usual aim of such studies: that of distinguishing the effects of the factors. A related consequence of inequality of sample size is that the various methods of computing SS statistics no longer yield identical results. Interpretations based on the diverse methods can differ profoundly, and the method to be preferred is often not obvious.

Missing cells

Data in which there are no observations for some combination(s) of treatments are said to have missing cells. This situation may be considered simply an extreme case of unequal sample size. In the realized data, the factors are not independent, and hence, inferences about the effects of the two factors are also not independent. As also with unequal sample size, results of the diverse methods of calculating ANOVA tables do not coincide. However, the case of missing cells contrasts with the case of unequal sample size in one important respect: whereas unequal sizes provide varying amounts of information for the different treatment combinations, *no* information is available for treatment combinations that are not observed at all. Special care must be taken in analyzing data having missing cells in order to take this absolute ignorance into account. As we will point out below, the methods recommended for analysis of the case of unequal sample size are not appropriate for analyzing data with missing cells. Moreover, general inferences about the effects of the factors are usually precluded entirely.

Missing responses

The above two types of imbalance refer to aspects of the design of the treatment combinations (a_i, b_j) and differ, albeit profoundly, only in the degree of imbalance. Alternatively, lack of balance may impinge on the response (y). The investigator may plan to measure several attributes of each organism in the experiment. Then, because analyses of the multiple measures on the same individuals are not independent, it is appropriate to conduct a multivariate ANOVA (MANOVA) to investigate the effect of the treatments on the responses jointly. While equal numbers of individuals may represent each cell of the experimental design, the design may be unbalanced in the sense that not all response variables were measured on each individual. One obvious reason for incompleteness of multivariate observations is mortality of individuals during the course of the experiment, or inability to locate individuals for a particular census. As noted above, care should be taken to account for systematic effects of the treatment combinations leading to missing values within the multivariate responses of particular observations.

Traditional Approaches to Lack of Balance

Because of the methodological difficulties that confront the investigator analyzing unbalanced data, efforts have traditionally focused on *imposing* balance. It is possible to delete observations chosen at random from those cells that have "extra" data, and then analyze a balanced subset of the data. Although statistically valid, this approach is undesirable since it does not use all available data, and is therefore likely to reduce the precision of estimates and the power of hypothesis tests. Moreover, different estimates are obtained depending on which data are deleted. In the particular case of multivariate analysis, several existing statistical packages eliminate all observations for which any of the measures is missing, resulting in a potentially drastic loss of available information. The MANOVA of such a reduced data set reveals the effects of the factors on *all* the response variables taken on only *part,* perhaps only a small part, of the design. An alternative is simply to restrict the analysis to a subset of the response variables measured on most or all of the individuals. Either procedure yields an analysis of less information than is available. The latter, however, permits the investigator to exercise control over the amount of data eliminated and the design of the data remaining.

As an alternative to eliminating data, one may fill in values estimated ("imputed") from the data, provided that only a few observations are missing, and then proceed with the standard ANOVA of the now-balanced data. Various methods have been proposed for obtaining such imputed values (e.g., substituting the cell mean for a missing observation, see Steele and Torrey 1980:209). Analysis of the resulting set of balanced data using standard methods, leads to correct estimates of the parameters, but biased tests of significance. More recently, novel methods of imputing missing values have been developed. These include, for the univariate case, Bartlett's ANCOVA method and the EM (expectation-maximization) algorithm of maximum likelihood estimation, related methods that produce asymptotically correct estimates and significance tests (Rubin 1976, Little and Rubin 1987). The EM algorithm can be applied to a diverse array of problems, including the case of missing observations in a multivariate response. This is an active area of research. Unfortunately, to our knowledge, general computer programs employing either Bartlett's ANCOVA or the EM algorithm for fixed effects ANOVA are not readily available.

Because the simple procedures for imposing balance are flawed, alternative methods for ANOVA were developed by Yates (1934) to take account of lack of balance. (See Herr [1986], for an interesting history of attitudes toward the diverse methods Yates proposed.) In the next section we review the methods currently in use. See also Potvin and Roff (1993) (this feature) for a discussion of the application of resampling methods as an alternative approach to ANOVA of unbalanced data.

ANOVA of Unbalanced Data

A number of methods for computing SS and testing hypotheses are available in existing statistical packages. Here we review them, using their designations from the SAS system as Type I–IV (Freund et al. 1986). Even for unbalanced data, they all produce the same ss and tests for the highest order interaction and the residual. For the main effects [and other interactions], however, the four methods may produce strikingly different results when the data are unbalanced. We continue to use the example of the two-way factorial design with interaction. In order to illustrate that results obtained by the various methods may differ profoundly, we consider a small example data set (Table 1, see Burdick and Herr [1980] and Milliken and Johnson [1984] for further examples). For the sake of discussion, we imagine an experiment to examine the effect on final plant height of removing conspecifics within a

TABLE 1. Example hypothetical set of unbalanced data for comparing results using Types I, II, and III analyses. Hypothetically defined initial size classes: 1 (small) and 2 (large). Hypothetical treatments: 0 = Control, no removal of neighboring plants; 1 = Removal of neighboring conspecifics within a given radius. Entries are observed responses (e.g., final plant height). Cell means and marginal means (as simple averages of the cell means) are given in brackets. These are the least squares means (Freund et al. 1986).

Initial size class	Treatment		
	0	1	
1	50	57	[62.25]
	57	71	
		85	
	[53.5]	[71.0]	
2	91	105	[108.87]
	94	120	
	102		
	110		
	[99.25]	[112.5]	
	[76.37]	[91.75]	

given distance of a target plant (Factor A). Thus, there is a removal treatment and a control to which the removal is to be compared. In the interest of determining whether there is also an effect of initial size on final height, and whether the effect of the removal treatment is independent of initial plant size, a second factor, initial size (Factor B), is crossed with the first. In our example, there is a maximum of four observations for each of the four cells, but one cell has two and another has three observations. We have deliberately kept this example small, so that it is easy to examine details of the data.

The Type I method sequentially fits each effect in the order that it appears in the model. It then computes SS accounted for by that effect. For example, assuming the model is given in the order specified in Eq. 2, RSS$_1$ is the residual SS from the model, $y_{ijk} = \mu + e_{ijk}$, and RSS$_2$ is the residual SS from the model, $y_{ijk} = \mu + a_i + e_{ijk}$. The reduction in the residual SS due to factor A adjusted for the mean, R($a \mid \mu$), is calculated as RSS$_1$ − RSS$_2$. (We here employ the widely used R($\cdot \mid \cdot$) notation somewhat reluctantly, in view of its inherent ambiguities, discussed by Searle et al. [1981]. However, these ambiguities do not affect the discussion in the general terms used here.) For each additional factor in the model, the Type I SS attributable to that new effect is calculated as the further reduction in residual SS due to that factor (i.e., for Factor B, R($b \mid \mu, a$), and for A × B, R($t \mid \mu, a, b$)). For a given ordering of the effects in the model, the SS for the different factors are independent of each other and sum to the total SS. However, SS and tests of significance for each factor depend on the order of factors in the model. In most instances, there will be little biological justification for this approach. Moreover, examination of the parametric expressions for Type I hypotheses (Speed et al. 1978, Milliken and Johnson 1984, Maxwell and Delaney 1989) reveals that Type I tests compare marginal means weighted by the cell sizes, and hence depend on realized sample sizes. Numerous authors have argued that details of the sampling scheme should not be involved in inference of general effects. It has further been noted that the Type I method tests hypotheses that can only be specified (exactly, in parametric terms) after the data have been collected. While this may be appropriate when it is of interest to compare the effects weighted by the frequencies with which the cells are represented, as might be the case in observational surveys (see example in Maxwell and Delaney 1989:274ff.), the statistical consensus is that Type I analyses of unbalanced data are not generally acceptable when inferences concern the effects themselves.

Referring to our example (Table 2), we can see that the Type I analysis indicates that there is no statistically significant effect of the removal treatment on final height. The effect of initial size on final height is highly significant. In addition, this analysis indicates no significant interaction. This interpretation of the data depends very heavily on the order of entry of the factors in the model. If we reverse the order, entering Size before Removal, we find that the effect of Size changes slightly (F = 40.17; P = .0004) and the effect of Removal is nearly significant (F = 5.52; P = .051). The statistics concerning the interaction are as shown for the original ordering.

The Type II method proceeds in a fashion similar to Type I, sequentially estimating effects and the SS associated with each effect as the amount that the residual SS are reduced by including it in the model.

TABLE 2. Analyses, using three different types of sums of squares (SS) in ANOVA, of the example of unbalanced data given in Table 1. Results are from SAS GLM procedure. P = probability of a greater F value.

Source	df	Type I SS	F	P
Treatment	1	35.3	0.33	.583
Size	1	4846.0	45.37	.0003
T × S	2	11.4	0.11	.753
Error	6	747.7		
Source	df	Type II SS	F	P
Treatment	1	590.2	5.52	.051
Size	1	4846.0	45.37	.0003
T × S	2	11.4	0.11	.753
Error	6	747.7		
Source	df	Type III SS	F	P
Treatment	1	597.2	5.59	.050
Size	1	4807.9	45.01	.0003
T × S	2	11.4	0.11	.753
Error	6	747.7		

While the Type I method computes RSS for a given effect based only on terms preceding it in the model, Type II computes RSS for all effects in the model that are at the same or lower level. For example, in the two-way factorial design, Type II ss for the main effects take account of all other main effects, rather than simply accounting for those entered earlier in the model. Interaction effects take account of all main effects and all other interaction effects at the same level. To compare Type I and Type II methods, the ss due to factor A equal R(a | μ) using Type I but R(a | μ, b) with Type II ss. Type II analysis is based on the assumption that the interaction is negligible or non-existent, and, in that case, is expected to achieve greater power for comparisons, relative to Type III (Burdick and Herr 1980). However, as with Type I analysis, Type II methods test parametric hypotheses that involve the sample sizes. They are therefore subject to the same criticisms.

For the example data, the results of the analysis using Type II ss differ quite markedly from those of the Type I method given in Table 2. These results indicate that the effect of the removal treatment is much greater than we inferred from our first Type I analysis. This difference is due to the fact that the Type II method assesses the additional effect of a given factor (here, Treatment) beyond the effects of the remaining factors at the same or lower level (here, Size). While some would not regard this result as "significant," most would agree that the removal effect is substantial and worthy of further consideration. Considering the effects of Size, the Type II results agree with those of the Type I analysis. For the second factor specified in the model, the Type I and Type II ss are identical, by the definition above, and this is also true for the interaction.

When all treatment combinations are observed, but the number of observations per cell varies, the Type III method provides the most readily interpretable tests of the null hypotheses of no main effect of Factor A and B (Speed et al. 1978, Milliken and Johnson 1984). The Type III ss for each main effect is the sum of the squared differences of *unweighted* marginal means, i.e., the least squares means (Table 1). The Type III ss do not, therefore, depend on details of the sampling structure in the data at hand. For this reason, they answer questions of general interest. They quantify the effect of a particular factor adjusted for all other factors in the model. For each factor, the Type III ss is the estimate that would be obtained from a Type I analysis of a model in which that factor appeared last. Thus, as with Type II analyses, Type III tests of the various factors do not depend on the particular order in the model.

Returning to the example data, we see that the results of the Type III analysis are in close agreement, in this case, with those from the Type II analysis (Table 2). The effect of the removal treatment is, however, now significant at the level, $P = .05$. It is worth noting that the Type II approach was developed specifically for cases in which the interaction is absent. In that particular case, Type II is more powerful than Type III. However, in our example the Type III test is more powerful. Thus, because finding that the interaction is not significant does not guarantee that there is none, it also does not guarantee that a Type II analysis is more powerful than Type III.

There is one aspect of the Type III method that might be considered a drawback. Given that the factors themselves are not generally independent, as a result of inequality of cell sizes, the Type III tests of the main effects of the two factors are not independent. Thus, it is not possible to judge the independent effect of each factor. This is a general consequence of multicollinearity, i.e., correlation among predictor variables, a widely recognized problem that plagues interpretation of multiple regression analyses (Draper and Smith 1981, Neter et al. 1983; Mitchell-Olds and Shaw [1987] consider this problem in the context of inferring selection on multiple characters). The same problem lurks behind unbalanced ANOVA. When the imbalance is slight, it is unlikely to seriously complicate interpretation of the test statistics. However, the consequence of severe imbalance, leading to extreme multicollinearity, is that the power of the design is reduced, and it is therefore quite possible to overlook an effect (i.e., judge it as not significant) when the effect truly exists. In the case of the two-way crossed design, one factor may indeed affect the response, while the second does not. If the two are highly correlated in the realized design, neither effect may appear significant according to the Type III tests, whereas a Type I or Type II analysis would indicate a significant effect of whichever factor appears first in the model. Although we join many statistical texts in recommending Type III tests over Type I and II when the data are unbalanced without missing cells, we urge particular caution in interpretation of failure to reject null hypotheses.

Beyond this general problem, Type III tests are invalid in the specific instance when there are missing cells (Milliken and Johnson 1984:185). In this particular case it is not possible, using a Type III analysis or any other approach, to test the standard main effect hypotheses of equality of marginal means, because the marginal means are not all estimable, by virtue of the absolute lack of information about the means of cells for which no data are available. Type III analyses can nevertheless be produced even when there are missing cells. Milliken and Johnson (1984:185), however, call them the "worst hypotheses to consider in this situation because there seems to be no reasonable way to interpret them."

Type IV ss are identical to Type III ss when all treatment combinations are observed. When some cells are empty, the Type IV approach takes into account the missing treatment combinations to develop particular testable hypotheses given the available data. Considering our example (Table 1) and assuming now that there are no observations of individuals in size class 1 exposed to treatment 1, it is possible to compare, for control plants, the effect of initial size on final height. It is also possible to compare, for plants initially in the large size class, the effect of the removal treatment compared to the control. The Type IV method makes these particular comparisons, using the appropriate cell means. It is thus not possible to make general inferences of the effect of the treatment or of initial size on final height. Given a two-way design in which there are more levels of each factor, it would be possible to develop various, potentially many, testable hypotheses for each main effect. Which hypotheses could be tested would depend on which cell means can be estimated from the data at hand. The Type IV method "chooses" a subset from among the testable hypotheses and provides ss and test statistics for these. Because there are other parametric hypotheses that correspond to each main effect, this should not be considered a complete analysis. (As the SAS output warns, "Other Type IV testable hypotheses exist which may yield different ss.") Milliken and Johnson (1984:187) point out that the set of hypotheses tested is arbitrary and may depend on the order that observations are entered in the data set. Thus, for a given analysis, the parametric hypotheses tested by default using Type IV methods *may* be of interest, but it is very possible that other Type IV hypotheses are of equal or greater interest. For data with missing cells, these authors recommend that the investigator carefully choose which combinations of cell means to compare (e.g., H_0: equality of particular combinations of estimable cell means). Particular hypotheses to test can be specified in SAS using options ESTIMATE or CONTRAST (Freund et al. 1986). Alternatively, the data can be analyzed as a one-way design, and tests can be carried out using confidence intervals of estimates of the cell means. We emphasize the care necessary to properly analyze data with missing cells by quoting Milliken and Johnson (1984:190): ". . . a good analysis of data with missing treatment combinations requires a great deal of thought. An experimenter or statistician cannot simply run a computer program on the data and then select numbers from that program to report in a paper. Unfortunately, this has been done and is being done by an extremely large number of experimenters and data analysts. We hope that anyone who has studied this chapter will never do it again."

Speed et al. (1978), Searle et al. (1981), Milliken and Johnson (1984:157, 190), and Maxwell and Delaney (1989:291) present more detailed comparisons of these four methods, including parametric expressions for the hypotheses tested by each and the correspondence between the designations I–IV in SAS and the methods available using SPSS and BMDP.

Conclusion

Analysis of unbalanced data in ecology will often present numerous difficulties, but some of these may be avoided with greater understanding of the methods and assumptions involved. Despite ample precedent for imposing balance on unbalanced data, either by eliminating values (and hence losing information) or by filling in missing values with cell means (leading to biased tests), it is generally preferable to use computational methods that are specifically designed for unbalanced data. Generally, tests based specifically on the cell means model (Eq. 1) may be more readily interpretable. However, the available statistical computing packages are designed for the effects model (Eq. 2). Milliken and Johnson (1984) show in detail how these packages can be used to obtain analyses in terms of the means model. Alternatively, careful analysis in terms of the effects model using Type III or Type IV methods can glean all the same information from the data at hand.

Further Readings and Applications

Texts we have found especially useful in clarifying when and how to apply the various methods include Milliken and Johnson (1984) and Maxwell and Delaney (1989). The three methods of analyzing unbalanced data, Types I through III, are all available in SAS: Proc GLM, BMDP-P4V, and SPSS: ANOVA and MANOVA. However, the default method differs among these packages. For example, Type II, which is not widely recommended, is the default method in SPSS-X ANOVA. The Type IV approach for analyzing data with missing cells is available in SAS and, with some limitation, BMDP-P4V (Milliken and Johnson 1984).

Acknowledgments

We thank Joe Felsenstein, Charlie Geyer, Alex Ramirez, and Bruce Walsh for helpful discussions, and Randy Mitchell, Gert Platenkamp, Robert Podolsky, Frank Shaw, Bruce Walsh, and Nick Waser for comments on the manuscript. R. G. Shaw wishes to especially thank D. S. Burdick for early guidance among these hazards. Shaw acknowledges support from NSF (BSR-8817756 and DIR-9112842). T. Mitchell-Olds was supported by grants from NSF (BSR-9021451 and BSR-9100397) and USDA (CRGP 88-37151-3958).

Literature Cited

Burdick, D. S., and D. G. Herr. 1980. Counterexamples in unbalanced two-way analysis of variance. Communications in Statistics, Part A—Theory and Methods **9**:231–241.

Draper, N. R., and H. Smith. 1981. Applied regression analysis. Second edition. Wiley, New York, New York, USA.

Freund, R. J., R. C. Littell, and P. C. Spector. 1986. SAS system for linear models. SAS Institute, Cary, North Carolina, USA.

Herr, D. G. 1986. On the history of ANOVA in unbalanced, factorial designs. American Statistician **40**:265–270.

Little, R. J. A., and D. B. Rubin. 1987. Statistical analysis with missing data. Wiley, New York, New York, USA.

Maxwell, S. E., and H. D. Delaney. 1989. Designing experiments and analyzing data: a model comparison perspective. Wadsworth, Belmont, California, USA.

Milliken, G. A., and D. E. Johnson. 1984. Analysis of messy data. Volume 1: designed experiments. Van Nostrand Reinhold, New York, New York, USA.

Mitchell-Olds, T., and R. G. Shaw. 1987. Regression analysis of natural selection: statistical inference and biological interpretation. Evolution **41**:1149–1161.

Neter, J., W. Wasserman, and M. H. Kutner. 1983. Applied linear regression models. Irwin, Homewood, Illinois, USA.

Potvin, C., and D. A. Roff. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? Ecology **74**:1617–1628.

Rubin, D. B. 1976. Noniterative least squares estimates, standard errors and F-tests for analyses of variance with missing data. Journal of the Royal Statistical Society **B38**: 270–274.

Searle, S. R. 1971. Linear models. Wiley, New York, New York, USA.

———. 1987. Linear models for unbalanced data. Wiley, New York, New York, USA.

Searle, S. R., F. M. Speed, and H. V. Henderson. 1981. Some computational and model equivalences in analyses of variance of unequal-subclass-numbers data. American Statistician **35**:16–33.

Shaw, R. G. 1986. Response to density in a wild population of the perennial herb, *Salvia lyrata*: variation among families. Evolution **40**:492–505.

———. 1987*a*. Maximum-likelihood approaches applied to quantitative genetics of natural populations. Evolution **41**: 812–826.

———. 1987*b*. Density-dependence in *Salvia lyrata*: experimental alteration of densities of established plants. Journal of Ecology **75**:1049–1063.

Speed, F. M., R. R. Hocking, and O. P. Hackney. 1978. Methods of analysis of linear models with unbalanced data. Journal of the American Statistical Society **73**:105–112.

Steele, R. G. D., and J. H. Torrey. 1980. Principles and procedures of statistics. Second edition. McGraw-Hill, New York, New York, USA.

Yates, F. 1934. The analysis of multiple classifications with unequal numbers in the different classes. Journal of the American Statistical Association **29**:51–66.