Statistics and Computing **13**: 163–167, 2003

ANOVA for unbalanced data: Use Type II instead of Type III sums of squares

ØYVIND LANGSRUD

MATFORSK, Osloveien 1, N-1430 Ås, Norway

Received June 2001 and accepted October 2002

Methods for analyzing unbalanced factorial designs can be traced back to Yates (1934). Today, most major statistical programs perform, by default, unbalanced ANOVA based on Type III sums of squares (Yates's weighted squares of means). As criticized by Nelder and Lane (1995), this analysis is founded on unrealistic models—models with interactions, but without all corresponding main effects. The Type II analysis (Yates's method of fitting constants) is usually not preferred because of the underlying assumption of no interactions. This argument is, however, also founded on unrealistic models. Furthermore, by considering the power of the two methods, it is clear that Type II is preferable.

Keywords: unbalanced factorial design, linear model, fixed effect, nonorthogonal, fitting constants, constraint

1. Introduction

There are three different classical approaches for computing sums of squares (SS) and testing hypotheses in ANOVA for unbalanced data. Using the designations from SAS, these methods are called Type I–III. Yates (1934) has described all three methods for two-way tables and during the years several authors have discussed these methods (see Speed, Hocking and Hackney (1978) and Herr (1986)). A frequent conclusion from the literature is that standard analysis should be based on Type III. As a result most major statistical programs perform the Type III analysis by default. Table 1 lists several well known statistical packages together with information about available and default SS types.

The Type I analysis corresponds to adding each effect sequentially to the model and it depends on how the model terms are ordered. Different orders may give quite different results. A small p-value for a certain factor is not necessarily a proof that the factor has an effect. Nevertheless, the method can be a powerful tool in the model building process. Below we discuss Type II and Type III which do not depend on how the model terms are ordered. Note, however, that the results from these methods can be obtained by running several Type I analyses using different orders.

In the Type III analysis each effect is adjusted for all other terms in the model. The result for each model term could be obtained by a Type I analysis where the actual term is the last one in the ordering. For this to be possible, a unique parametric form of the full model has to be specified. Since the ordinary model with main effects and all interactions is overparameterized, constraints on the parameters are needed. The Type III analysis is based on the usual constraints where the parameters for each term add to zero when summed over any subscript (see (3) below). Nelder (1977, 1994) and Nelder and Lane (1995) have criticized the Type III tests because they correspond to uninteresting hypotheses—namely tests of main effects in the presence of interactions. In the discussions of Nelder (1977, 1994) several authors agree with Nelder. Kempthorne (1975) says that "*The testing of main effects in the presence of interaction, without additional input, is an exercise in fatuity.*"

In the Type II analysis each effect is adjusted for all other terms except terms that "contain" the effect being tested. For example in a three-way table (A, B and C) the main effect of factor A is not adjusted for the interactions *AB*, *AC* and *ABC*. And the two-factor interactions are not adjusted for *ABC*. As mentioned above, the Type II results can be obtained by using several Type I analyses. In the two-way case (A and B), only two Type I analyses are needed; the order A-B-AB (Type II results for B and AB) and the order B-A-AB (Type II results for A and AB). The Type II tests do not make use of constraints on the parameters. That is, alternative model specifications produce identical results. The Type II method is usually not preferred because it is based on the assumption that the interactions are negligible or non-existent. However, Overall and Spiegel (1969)

Table 1. *Overview of available SS types in 12 statistical packages: A = available and D = default. Note that Type I SS are often called sequential SS and Type III SS are sometimes called partial or adjusted SS*

Program name and version	Type I	Type II	Type III
GENSTAT 6[a]	D	A	
JMP 5[b]	A		D
MINITAB 13	A		D
R 1.5[c]	D		
SAS (SAS/STAT) 8[d]	D	A	D
S-PLUS 6[e]	D		A
SPSS 11[d]	A	A	D
STATA 7	A		D
STATGRAPHICS Plus 5	A		D
STATISTICA 6[d,f]	A	A	D
SYSTAT 10			D
UNISTAT 5[g]	A	A	D

[a]Type II is referred to as the conditional test. The so-called marginal test is another available method.
[b]The default is the so-called effective hypothesis (Statistica's Type VI) which is equivalent to Types III and IV when no cells are missing.
[c]Types II and III are available through the car package.
[d]Type IV is available.
[e]Type II is available through the car package.
[f]The so-called Types V and VI (effective hypothesis) are available.
[g]Types I-III are called, respectively, the classic experimental approach, the hierarchical approach and the regression approach.

consider the Type II method "*most appropriate for experimental research which is viewed in the context of traditional analysis of variance*". Below we will discuss the two methods, Type II and Type III, further.

2. Two-way ANOVA

Consider a factorial design with two treatment factors; A with a levels and B with b levels. Based on the cell means the model can be written as

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

($i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; $k = 1, 2, \ldots, n_{ij}$), where y_{ijk} is the kth observation in the cell defined by the ith level of A and the jth level of B, μ_{ij} is the cell mean (expected value) and ε_{ijk} is the residual error (independent normally distributed with zero means). Usually this model is written using another parameterization:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}. \tag{2}$$

The parameter μ is a general mean, α_i is the A effect, β_j is the B effect and γ_{ij} is the interaction effect. This model is overparameterized and constraints are needed to obtain unique estimates. It is customary to impose the following restrictions on the parameters:

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \gamma_{ij} = 0 \,\forall\, j, \quad \sum_{j=1}^{b} \gamma_{ij} = 0 \,\forall\, i. \tag{3}$$

Alternatively, other restrictions may be used.

The aim of doing a hypothesis test can be formulated at different levels:

Aim 1: To do inference about parameters.
Aim 2: To guide the model selection.
Aim 3: To answer specific questions.

Parameters are necessary for describing and using models. It is, however, meaningless to test hypotheses about parameters without foundation in Aim 2 or Aim 3. Following Nelder (1977), the specification and the fitting of models (Aim 2) is called the *smoothing phase*. Beyond this is the *prediction phase*, when predictions are made and specific questions are answered (Aim 3). However, several questions could have been answered during the *smoothing phase*, e.g. whether a specific factor has any effect. Indeed, *to obtain answers* (Aim 3) is always the main goal. We use parametric models as powerful tools.

2.1. *Model selection*

We now consider the second aim. As argued by the critics of the Type III analysis, there is no doubt that a model including interaction but without both main effects is unrealistic. If the interaction is needed, this means that factor A has an effect. In a model without main factor A (all $\alpha_i = 0$) the effects of A summed over the levels of B are exactly zero (according to the constraints (3)). There is, however, no reason to make such an assumption. Nelder (1994) writes: "*Now since the presence of the interaction places no restrictions at all on how A varies as B changes (and vice versa), we ought to be very surprised if either margin were null. Why should it be?*" Accordingly, we have only five realistic model alternatives:

Model 1: No effect of A and no effect of B.
Model 2: Only effect of A.
Model 3: Only effect of B.
Model 4: The additive model (without interaction).
Model 5: The full model (with interaction).

The objective of the hypothesis testing is now to choose among these five alternatives. We will denote the different SS's by: SS_A^{II}, SS_A^{III}, SS_B^{II}, SS_B^{III}, SS_{AB}, and SS_{ERROR}. The upper subscript distinguishes between Type II and Type III when these are different.

The Type III test is based on the model equation for the full model (2) with the restrictions (3) and the hypothesis being tested is simply and directly

$$H_0\colon \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0. \tag{4}$$

The corresponding Type III F-statistic is

$$F = \frac{\mathrm{SS}_A^{\mathrm{III}}/\mathrm{df}_A}{\mathrm{SS}_{\mathrm{ERROR}}/\mathrm{df}_{\mathrm{ERROR}}} \tag{5}$$

where df is degrees of freedom. This test statistic is strange because $\mathrm{SS}_A^{\mathrm{III}}$ is adjusted for the interaction. This adjustment follows from the fact that the interaction (the γ_{ij}-parameters) is a part of the H_0-model. This H_0-model is, however, not one of the realistic models listed above. The model under the null hypothesis has to be Model 1 or Model 3. Testing under Model 1 corresponds to sequential model building (Type I analysis). We will here concentrate on Model 3 as the null model. When Model 4 is the alternative model the hypothesis (4) can be tested according to classical regression theory and under the assumption of no interaction there is no difference between Type II and Type III. The F-statistic can be expressed as

$$F = \frac{\mathrm{SS}_A^{\mathrm{II}}/\mathrm{df}_A}{(\mathrm{SS}_{\mathrm{ERROR}} + \mathrm{SS}_{AB})/(\mathrm{df}_{\mathrm{ERROR}} + \mathrm{df}_{AB})} \tag{6}$$

which means that the Type II SS is independent of a possible assumption of no interaction. When we may have interaction, statistic (6) is not optimal. Model 4 is the main alternative since the interaction will be tested separately. But the possibility of the wider alternative, Model 5, should not be neglected. When using the F-statistic in (6) a possible interaction effect can reduce the power since SS_{AB} is a part of the error term. This is avoided by removing SS_{AB} from the denominator and the statistic becomes

$$F = \frac{\mathrm{SS}_A^{\mathrm{II}}/\mathrm{df}_A}{\mathrm{SS}_{\mathrm{ERROR}}/\mathrm{df}_{\mathrm{ERROR}}} \tag{7}$$

which is the Type II test statistic. If there is no interaction this test will have less power than (6) since the error term is based on fewer degrees of freedom. However, in cases were $\mathrm{df}_{\mathrm{ERROR}}$ is not very small there is only a slight difference in power. In ANOVA analysis one will always have in mind that there should be "enough" degrees of freedom for error.

For the interaction test there is no difference between the two analyses. To summarize the discussion so far, it would be natural to replace the Type III tests by the Type II tests when the objective is model selection.

What about the assumption of negligible interactions underlying the Type II analysis? This assumption is the reason why the Type II tests are not widely used. However, this argument against the Type II analysis can be answered by an argument against the Type III analysis. The analysis should only be based on realistic models. If the interaction is present, hypotheses about the main effects are uninteresting. The results from the Type II analysis can be interpreted as follows: If a main effect is found to be significant, this result is correct if there is no interaction. If the interaction is present, both main effects will also be present. In any case, the statement about a significant main effect is correct.

Even if the reasoning underlying the calculation of the Type III test statistic is questionable, a possible significant Type III result is not wrong. To choose among a Type II test or a Type III test is in reality a question of choosing the most powerful test. When there is no interaction, there is no doubt that the Type II test is most powerful. Possible support for Type III must be based on cases where the interaction is present. In such cases we have two tests that may discover that A is affecting the response, namely the test of A and the test of AB. One could calculate a joint SS for the two terms, say SS_{A+AB}. As opposed to the Type III tests, the Type II tests are based on an orthogonal decomposition of SS_{A+AB}. Intuitively, without more information, an orthogonal decomposition is preferable. The specific type of decomposition is also intuitively appealing. The main effect component, $\mathrm{SS}_A^{\mathrm{II}}$, is the variation that can be explained by an additive model and this component tends to be most important. But the actual power depends on the structure of the interaction and it can not be guaranteed that Type II is most powerful in all cases (Shaw and Mitchell-Olds, 1993). Lewsey, Gardiner and Gettinby (1997, 2001) have performed some limited simulation studies of the power and their conclusion is that the Type II method is, on average, more powerful than the Type III method. To summarize this power discussion, it is clear that Type II performs better than Type III.

Above we have argued against testing the main effect toward the full model as the usual alternative. However, much of the historical discussion has focused on "What is the hypothesis being tested in terms of parameters under the full model?". As mentioned earlier, Type III is testing the correct hypothesis (4). Under the full model the Type II tests would test linear combinations of parameters that involve interactions (γ_{ij}). Following Speed, Hocking and Hackney (1978) the Type II and Type III hypotheses can be formulated according to the cell means model (1) as

$$H_0^{\mathrm{II}}\colon \sum_{j=1}^{b} n_{ij}\mu_{ij} = \sum_{j=1}^{b} n_{ij}\left(\sum_{i'=1}^{a} n_{i'j}\mu_{i'j} \Big/ n_{\cdot j}\right) \quad \forall\, i \tag{8}$$

$$H_0^{\mathrm{III}}\colon \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \cdots = \bar{\mu}_{a\cdot} \tag{9}$$

where $\bar{\mu}_{i\cdot} = \sum_{j=1}^{b} \mu_{ij}/b$ (for $i = 1, 2, \ldots, a$) and $n_{\cdot j} = \sum_{i=1}^{a} n_{ij}$. From this viewpoint the classical argument is that the Type II tests are difficult to justify because the hypotheses being tested depend on the cell frequencies (n_{ij}-values). However, in our discussion so far, the aim of our hypothesis test is model selection (Aim 2). For this purpose the tests do not have to be interpreted as a test of linear combinations of parameters under the full model. It seems that the logical ordering of Aim 1–3 has been reversed in the sense that inference about parameters (under the full model) is the main purpose.

It is possible to change the constraints described in (3). In that case the corresponding Type III tests would also change. The Type II tests are, however, not dependent on the constraints. Nelder (1994) argues against the use of constraints and he says that "*such constraints are not an intrinsic part of the model*". Furthermore, he claims that the constraints have confusing

consequences and they are the reason why uninteresting hypotheses (Type III) are performed.

2.2. *Tests for specific questions*

So far we have not treated Aim 3 explicitly. As mentioned earlier some questions could have been answered during the model selection process. Other questions can be answered by using and interpreting the final fitted model. Some of these questions might be answered by performing hypothesis tests. To illustrate this topic we consider an example originally described in Elston and Bush (1964):

> "Suppose we have two different feeding treatments as the two levels of factor A and three different breeds of cows as the three levels of factor B, and the response that is measured on each cow is some index of its productivity. Suppose further that one of the feeding treatments is a standard feed and that the other is a new feed; the question is whether or not to recommend the new feed in a certain area where 80% of the cows are of one breed, 5% are of another and 15% are of a third breed.... In this simple case the hypothesis reduces to H_0: $16\mu_{11} + \mu_{12} + 3\mu_{13} = 16\mu_{21} + \mu_{22} + 3\mu_{23}$. Furthermore, even if we are fairly confident that the interaction effects are small, it would be preferable to test a hypothesis of this form rather than arbitrarily to assume a model with no interaction term in it."

In this example a specific hypothesis is tested under the full model with interactions. The hypothesis test is not related to model selection. In fact there is no reason to believe in the null hypothesis as a plausible model alternative. The test is performed as tool for answering the question: "Can we from this data material conclude that a specific feed is preferable?" In a slightly modified version of this example there is one third of each breed. In that case all weights in the expression for the hypothesis would be one (instead of 16, 1 and 3). This means that the actual hypothesis is precisely the Type III hypothesis (9). In this situation the Type III hypothesis is justified and it is more natural than a Type II hypothesis. In the discussion of Nelder (1977), similar examples are described by Tukey (p. 72) and Frane and Jennrich (p. 73).

Based on such examples we will not change the conclusion that Type II is preferable. In the *prediction phase* one may ask questions that correspond to specific hypotheses. In some cases these hypotheses may be of Type III. But this argument can not justify that the automatic ANOVA output in computer programs should be of Type III. A specific hypothesis corresponding to a specific question can be tested by a specific user command (regardless if it corresponds to Type III or not). Usually, the automatic ANOVA output is used to test null hypotheses that are plausible model alternatives; one wants to choose among different model alternatives or to find out whether certain model components are significant.

2.3. *When factor A has two levels*

The example above deals with the interesting special case of two-way ANOVA where factors A has two levels (two different treatments). Unbalanced ANOVA for this case was originally introduced by Brandt (1933) and important comments were given by Yates (1934). The treatment effects within each level of factor B (sub-class effects) are defined by: $\text{effect}_j = \mu_{2j} - \mu_{1j}$. The overall treatment effect may be obtained as a weighted average of these sub-class effects and this overall effect can be subjected to a hypothesis test. In the example above one may use the weights 80, 5 and 15% (which are equivalent to 16, 1 and 3). The corresponding hypothesis test is exactly the one mentioned in the example. With all weights equal, the t-test for the overall effect is equivalent to the F-test based on SS_A^{III}. On the other hand, SS_A^{II} corresponds to weights according to the overall effect that has minimum variance.

Unbalanced analysis for the two-treatment case is of special importance for multicenter clinical trials. In that context, according to Senn (2000a), Type II is more common than Type III outside pharmaceutical industry. Senn (1998, 2000a, 2000b) and Gallo (2000) argue against Type III. To quote Senn (1998): "*My own view is that, although the type III philosophy seems plausible at first, it is untenable. It leads to paradoxes. For example, given two centres, a large and a small centre, unless the small centre is at least one-third the size of the large centre, the type III treatment estimate will have a larger variance than that based on the large centre alone. Thus more information is worse than less.*" And one of the comments in Senn (2000b) is: "*If we wish to show that the treatment can be effective, the natural null hypothesis is that it is identically 0 in every centre. The only relevant consideration therefore is what is the most effective combination?*" This statement coincides with the discussion above where Model 3 is the null hypothesis. In addition it is clear that we can show that the treatment has effect without focusing on a main effect defined according to specific constraints. This point is also important beyond the two-treatment case. The result from a Type II test (as well as a Type III test) can be used to prove that a certain factor has effect. A significant main effect means that the factor has significantly influenced the response variable.

3. General ANOVA

In the general case, the criticism of the Type III tests does also apply to two factor interactions and higher order terms. If for example a three factor interaction, say ABC, is needed, there is no reason to assume a model without all the two factor interactions, AB, AC and BC. A test of AB under the presence of ABC is meaningless. Therefore, there is no reason to adjust the SS for ABC, as is done in the computation of $\text{SS}_{AB}^{\text{III}}$. The SS's computed in the Type II analysis are not in conflict with the realistic models and there is no need to specify constraints on parameters. The reasoning underlying the test statistics is similar to the reasoning for two-way

designs. In each test three different model alternatives are considered:

The H_0 model: The model without the term being tested and without higher order terms that "involve" that term.
The narrow H_1 model: The term being tested is added to the H_0 model.
The wide H_1 model: This model contains all terms that are selected for the actual ANOVA analysis.

The wide H_1 model is the one that is specified by the user before the ANOVA analysis. This model is not necessarily the full model with all main effects and their interactions of all orders. The H_0 model is obtained by removing terms as described from the wide H_1 model. Based on this hierarchy of models, the total sum of squares can be decomposed into four orthogonal components: SS_{H0}, SS^{II}_{TERM}, SS_{EXTRA} and SS_{ERROR}. The Type II SS for the term being tested is SS^{II}_{TERM}, and SS_{EXTRA} is obtained by going from the narrow to the wide H_1 model. The test is directed towards the narrow H_1 alternative and therefore only SS^{II}_{TERM} goes into the numerator of the F statistic. Since there is a possibility of the wider alternative, SS_{EXTRA} is not incorporated into the denominator of the F statistic. When the narrow H_1 alternative is true, there is no doubt that the Type II test has more power than the corresponding Type III test. When this narrow model does not hold it is impossible to say that one test is always most powerful. However, the actual test and the tests for the extra terms in the wide model should be viewed together. All these tests may prove that our specific term is related to the response. For example, in the three-way case, a clearly significant three-factor interaction (*ABC*) means that we can conclude that there is interaction between factor A and factor B (as well as AC and BC). This means that the AB-phenomenon is tested by two hypotheses (*AB* and *ABC*) and in the Type II case these two tests are independent (SS according to two orthogonal components). Therefore the Type II test seems more powerful and in addition it is advantageous that the Type II test of *AB* is precisely a test related to the phenomenon that can be explained by a second order model (SS^{II}_{AB} is independent of eventual assumption of no three-factor interaction). We now turn back to the general case. In both Type II and Type III analyses all tests for the additional terms are based on components of SS_{EXTRA}. Since SS^{II}_{TERM} is orthogonal to SS_{EXTRA} the Type II test is preferable also when the narrow alternative does not hold.

4. Concluding remarks

The discussion above is limited to a comparison of Type III tests versus Type II tests and the conclusion is that Type II is preferable. This is not to say that a single Type II analysis can handle all problems. In many cases one needs to perform several analyses with different models and well-founded Type I tests are often very useful. The main message is that there are very strong reasons to prefer Type II over Type III and we recommend Type II as a better default choice than Type III.

Acknowledgments

I would like to thank the associate editor, the referees and editor Wayne Oldford for their insightful remarks and constructive suggestions, which led to a substantial improvement in the paper. I am also grateful to Per Lea for his helpful comments.

References

Brandt A.E. 1933. The analysis of variance in a $2 \times s$ table with disproportionate frequencies. Journal of the American Statistical Association 28: 164–173.

Elston R.C. and Bush N. 1964. The hypotheses that can be tested when there are interactions in an analysis of variance model. Biometrics 20: 681–698.

Gallo P.P. 2000. Center-weighting issues in multicenter clinical trials. Journal of Biopharmaceutical Statistics. 10: 145–163.

Herr D.G. 1986. On the history of ANOVA in unbalanced, factorial designs: The first 30 years. The American Statistician 40: 265–270.

Kempthorne O. 1975. Fixed and mixed models in the analysis of variance. Biometrics 38: 613–621.

Lewsey J.D., Gardiner W.P., and Gettinby G. 1997. A study of simple unbalanced factorial designs that use type II and type III sums of squares. Communications in Statistics—Simulation and Computation 26: 1315–1328.

Lewsey J.D., Gardiner W.P., and Gettinby G. 2001. A study of type II and type III power for testing hypotheses from unbalanced factorial designs. Communications in Statistics—Simulation and Computation 30: 597–609.

Nelder J.A. 1977. A reformulation of linear models (with discussion). Journal of the Royal Statistical Society Series A 140: 48–77.

Nelder J.A. 1994. The statistics of linear models: Back to basics (with discussion in vol. 5 (1995) 84–111). Statistics and Computing 4: 221–234.

Nelder J.A. and Lane P.W. 1995. The computer analysis of factorial experiments: In memoriam—Frank Yates. The American Statistician 49: 382–385.

Overall J.E. and Spiegel D.K. 1969. Concerning least squares analysis of experimental data. Psychological Bulletin. 72: 311–322.

Senn S. 1998. Some controversies in planning and analysing multicentre trials. Statistics in Medicine 17: 1753–1765.

Senn S. 2000a. The many modes of meta. Drug Information Journal 34: 535–549.

Senn S. 2000b. Consensus and controversy in pharmaceutical statistics. Journal of the Royal Statistical Society Series D—The Statistician 49: 135–156.

Shaw R.G. and Mitchell-Olds T. 1993. ANOVA for unbalanced data: An overview. Ecology 74: 1638–1645.

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

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