Unequal Weights in the Two-Way Analysis of Variance

by

Myra L. Samuels
Purdue University

Technical Report #85-31

Department of Statistics
Purdue University

November 1985

Unequal Weights in the Two-Way Analysis of Variance

by
Myra L. Samuels
Purdue University

ABSTRACT

In a two-factor analysis of variance, the population marginal means for Factor A are usually defined by assigning equal weight to each level of Factor B. It is argued that the advantages of using unequal weights have been under-appreciated. Unequal weights may lead to more natural interpretation of results. Furthermore, analysis using unequal weights can yield considerable gains in efficiency when the cell frequencies are unbalanced.

1. INTRODUCTION

Consider a two-way fixed-effects analysis of variance with two factors, A and B, where factor A is of primary interest and factor B is regarded as a nuisance factor. A commonly used analysis strategy is to test for interaction between factors A and B and, if no interaction is detected, to proceed to test for the A effect within the full model (that is, the linear model including the $A \times B$ interaction term). Two facts concerning this analysis strategy are:

Fact 1. The strategy requires that the data analyst specify weights to be used in the analysis.

Fact 2. The use of equal weights can be seriously inefficient if the cell frequencies are highly unbalanced.

Fact 1 is recognized in the literature (see Steinhorst 1982 and references therein), but Fact 2 has been neglected. It would appear that the use of unequal weights is widely regarded as an obscure tool to be applied only in special situations. Textbooks give the subject little if any space. Statistical computing packages use equal weights by default and some do not permit the use of unequal weights. Steinhorst (1982) points out that unequal weights may sometimes be appropriate; he does not discuss the question of efficiency. Further, Steinhorst expresses the opinion that the choice of weights should not depend on the observed cell frequencies.

The purpose of this paper is to indicate the advantages—increased efficiency and better

interpretability—of using appropriate unequal weights, and in addition to argue that in certain situations the choice of weights *should* depend on the observed cell frequencies.

2. EXAMPLES

To fix ideas, consider a medical experiment to compare two treatments for a particular disease. Patients are to be randomly allocated to two treatment groups (Factor A), and a response variable Y will be measured for each patient. Sex (Factor B) is considered a nuisance factor. The following are two typical scenarios which lead to a two-way analysis of highly unbalanced data.

Scenario 1. The investigators know that the male:female ratio is 9:1 in the population of interest (patients with the disease); consequently the randomization is stratified by sex to reflect this sex ratio.

Scenario 2. Patients are randomized to the two treatments without regard to sex. The statistician first analyzes the data ignoring sex. However, he notices that (either by chance or perhaps because of differential dropout rates) the sex distribution differs in the two treatment groups; to be on the safe side, he also tries analyses with sex in the model.

We now provide the above scenarios with artificial data sets to show how the use of equal weights can lead to seemingly paradoxical results.

Example 1. Suppose that the investigation of Scenario 1 includes 100 patients and yields a within-cell mean square $MSW = 1600$ and cell means as follows (cell frequencies are shown in parentheses):

		(Factor B) Sex	
		Male	Female
(Factor A) Treatment	1	335 (45)	325 (5)
	2	310 (45)	300 (5)

The statistician who tests for the effect of Factor A using the "usual" method (for instance, Type IV SS's in SAS GLM) would obtain the following significance probabilities:

$P = .0022$ in the additive model

$P = .0638$ in the full model

Because the cell means show no interaction at all, the estimated treatment difference is the same (+25) in both analyses; and yet the P-value differs by more than an order of mag-

nitude. The statistician might be puzzled that testing without the additivity assumption is so very much weaker; after all, it only "costs" a single *df* for error.

□

Example 2. Suppose that the investigation of Scenario 2 yields the following data:

(Factor A) Treatment	(Factor B) Sex: Male	Female
1	325 (45)	325 (5)
2	300 (35)	300 (5)

Again suppose that the full-model analysis is the "usual" one with equal weights. If the within-cell mean-square is MSW=1600, the following significance probabilities would be obtained:

$P = .0021$ ignoring sex
$P = .0022$ in the additive model
$P = .0289$ in the full model

Because the observed cell means do not show any sex effect at all, the estimated treatment difference is the same (+25) for all three analyses; the inclusion of sex in the model has only a mild effect if it is included additively, yet it is extremely costly if it is included in the full model.

□

In the next section, the puzzles confronting the statisticians of Examples 1 and 2 will be clarified by showing, not that the full model is somehow wrong for them, but that the "usual" version of the full model is seriously inappropriate for their purposes.

3. HOW WEIGHTS AFFECT THE ANOVA

Let the random variable Y_{ijk} represent the kth observation in the (i,j)th cell; assume that

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad i = 1,\ldots,a; \quad j = 1,\ldots,b; \quad k = 1,\ldots,n_{ij}.$$

where the $\{\varepsilon_{ijk}\}$ are independent normal random variables with mean zero and common variance σ^2. It is well-known (see, for example, Searle, Speed, and Henderson, 1981) that if the full model is written as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \tag{1}$$

then the parameters $\{\alpha_i\}$, $\{\beta_j\}$, and $\{\gamma_{ij}\}$ are not estimable unless constrained by side conditions. Suppose that we write the side conditions as follows:

$$\sum_i v_i \alpha_i = 0, \qquad \sum_j w_j \beta_j = 0,$$

$$\sum_i v_i \gamma_{ij} = 0, \qquad \sum_j w_j \gamma_{ij} = 0,$$

where $\{v_i\}$ and $\{w_j\}$ are positive weights satisfying

$$\sum_i v_i = 1; \qquad \sum_j w_j = 1.$$

The null hypothesis of no effect of Factor A can be stated as

$$H_A: \quad \alpha_1 = \alpha_2 = \ldots = \alpha_a.$$

Let us consider a more natural parameterization; let

$$A_i = \sum_j w_j \mu_{ij}, \qquad i = 1, \ldots, a.$$

The parameters $\{A_i\}$ might be called "weighted population marginal means" (WPMM's), after Searle, Speed, and Milliken (1980), who proposed the term "population marginal mean" for the case of equal weights. The least-squares estimate of A_i is

$$\hat{A}_i = \sum_j w_{ij} \overline{Y}_{ij},$$

where $\overline{Y}_{ij}$ is the observed mean in the (i,j)th cell.

The null hypothesis H_A can be restated in terms of the WPMM's as

$$H_A: \quad A_1 = A_2 = \ldots = A_a. \tag{2}$$

Note that the weights $\{v_i\}$ have no impact on H_A; but changing the weights $\{w_j\}$ changes the meaning of H_A. This formulation of H_A, which was proposed by Scheffé (1959), is a very natural one. The hypothesis H_A asserts that Factor A has no effect, after "adjusting for", or "taking account of", Factor B. *It is crucial for this interpretation that the weights $\{w_j\}$ do not depend on i.* (If the weights depend on i—as they may, for example, in the hypotheses H_2 and H_3 identified by Speed, Hocking and Hackney (1978)-or the hypothesis tested by the WEIGHTS BETWEEN ARE SIZES statement in BMDP4V— then the WPPM's are in no sense "standardized" with respect to Factor B.) The "usual" analysis (for instance, that implicit in SAS GLM Type IV SS's) uses equal weights

$$w_j = b^{-1}, \qquad j = 1, \ldots, b.$$

In a purely additive situation the weights are irrelevant. Specifically, let us state the hypothesis of no interaction as

$$H_{AB}: \quad \gamma_{ij} = 0, \qquad i = 1, \ldots, a; \qquad j = 1, \ldots, b.$$

It is easy to show that if H_{AB} is true for one set of weights $\{v_i\}$ and $\{w_j\}$, then it will be true for any set; also, the usual F-test for H_{AB} is independent of the weights. If H_{AB} is true, then the values of contrasts in the A_i, and consequently the truth or falsity of H_A, do not depend on the choice of weights.

To test H_A in the full model, the statistician must choose weights $\{w_j\}$ (or opt for the "usual" equal weights). In discussing the choice of weights, we still consider the following two F-statistics. The full-model F-statistic is of the form

$$F = \frac{MSA}{MSW} \tag{3}$$

where MSW is the within-cell mean square (with $df = \Sigma\Sigma n_{ij} - ab$), and MSA is the mean square for Factor A within the full model (with $df = a - 1$); the value of MSA depends on the weights $\{w_j\}$, but the value of MSW does not. The additive-model F-statistic is

$$F = \frac{MSA^*}{MSE} \tag{4}$$

where MSE is the residual mean square (with $df = \Sigma\Sigma n_{ij} - a - b + 1$) after fitting the additive model, and MSA* is the mean square (with $df = a - 1$) for Factor A within the additive model.

Suppose that in fact interaction is absent, that is, H_{AB} is true; in this case the hypothesis H_A does not depend on the weights, and one may ask which choice of weights in (3) provides the most powerful test of H_A. The answer is particularly simple in the important special case of proportionate cell frequencies, that is, if the cell frequencies $\{n_{ij}\}$ satisfy the conditions

$$n_{ij} = r_j n_{i+}, \qquad i = 1, \ldots, a; \qquad j = 1, \ldots, b, \tag{5}$$

where $n_{i+} = \sum_j n_{ij}$. (Proofs are in the Appendix.)

Proposition 1 Assume that (5) holds and that H_{AB} is true. Then the power of the F-test (3) of the hypothesis H_A is maximized for weights proportional to the cell frequencies, namely,

$$w_j = r_j. \tag{6}$$

Further, the standard error of any estimated contrast in the $\{A_i\}$ is minimized by use of the weights (6). The relative efficiency of the analysis using weights (6), relative to using equal weights, is:

$$\text{RE(optimal wts : equal wts)} = \frac{1}{b^2}\sum_j \left(\frac{1}{r_j}\right). \qquad (7)$$

□

Of course, if interaction is absent one could use the additive-model F-test based on (4) which is more powerful than any test based on (3). The difference between these two approaches is not very great if the cell frequencies are proportionate, as the following proposition shows.

Proposition 2 If the cell frequencies satisfy (5), then the values of MSA using the optimal weights (6) is equal to the value of MSA*. Similarly, in calculating a confidence interval for any contrast in the $\{A_i\}$, the results from an additive analysis and from a full-model analysis with weights (6) would differ only with respect to the estimate of σ and the associated *df*.

□

Example 1, revisited. In light of Propositions 1 and 2 let us reconsider the position of the statistician in Example 1. His full-model P-value was inflated because it was based on equal weights; the P-value for the F-test based on the optimal weights $(w_1 = .9, w_2 = .1)$ is $P = .0024$; the relative efficiency of this analysis, from (7), is $4^{-1}(.9^{-1} + .1^{-1}) = 2.8$; thus the equal-weights analysis wastes nearly 2/3 of the information in the data.

□

Intuitively, the reason that the equal-weights analysis is so inefficient in Example 1 is as follows: The cell means for the males and for the females are given equally important estimation jobs, but the sampling error is large in the female samples (because the n's are small), so that the entire estimation is relatively imprecise. When optimal weights are used, the performance demanded of each cell mean is matched to its capability.

Example 2, revisited. The cell frequencies in Example 2 are not proportionate. The statistician might reasonably (as we will argue below) choose weights $w_1 = .8$ and $w_2 = .2$ which agree with the observed marginal cell frequencies. A full model-analysis with these weights yields $P = .0034$, which is comparable to the additive analysis and to the analysis ignoring sex.

□

4. CHOICE OF ANALYSIS STRATEGIES

4.1 Proportionate Cell Frequencies

The typical situation giving rise to proportionate cell frequencies would be a stratified design like that of Scenario 1. In choosing a strategy for analyzing the effect of Factor A in the design of Scenario 1, the statistician might reason as follows:

If in fact interaction is absent (H_{AB} is true), then the analysis in the additive model is most efficient, but the analysis in the full model using the optimal 9:1 weights is nearly as efficient, differing only in the loss of 1 *df* for error.

If in fact interaction is present (H_{AB} is false), 9:1 weights make sense because with these weights the WPPM's A_1 and A_2 represent the actual population mean responses to treatments 1 and 2. Further, it can be argued that the existence of interactions does not seriously compromise the interpretation of the WPMM's: after all, in any study, treatment means are averaged over various extraneous variables, and perhaps sex should be viewed in that way. In addition, if H_{AB} is false, testing H_A with the additive-model statistic (4) would be inefficient because the denominator mean square would be inflated by the interactions.

Reasoning in this way, the statistician might adopt, as the best overall analysis strategy, the full-model analysis using optimal weights. Of course, the analysis of Factor A would naturally be supplemented by a test for interaction and possibly also for Factor B.

4.2 Non-proportionate Cell Frequencies

In many situations, including that of Scenario 2 as well as many observational studies, the cell frequencies might be more or less disproportionate, either by chance or because of differential dropout rates or other distortions with respect to Factor B. The question of "adjusting" for B in the analysis might arise from a desire for increased power and/or because B is a potential confounding variable. For instance, in Scenario 2, if the sex distribution in the two treatment groups is very different (in other words, if the cell frequencies are greatly disproportionate), then one might want to compare "sex-adjusted" treatment means (that is, $\hat{A}_i$'s) in order to eliminate bias.

For "adjustment" in the full model, considerations of interpretability of the WPPM's do not necessarily lead to weights which are optimal with respect to efficiency. Interpretability presumably should be the more important consideration; for instance, in Scenario 2, the sex-adjustment ideally should use weights which reflect the population sex ratio.

In some situations the best source of information about the population distribution of Factor B might be the current study. In such a case it appears eminently reasonable to use the information provided by the cell frequencies; often it would be natural to choose weights proportional to the marginal cell frequencies. Of course, this approach involves an approximation, in that the sampling error in the weights is ignored, and thus would not be appropriate if the marginal cell frequencies were small. (Weights proportional to marginal cell frequencies are not optimal with respect to efficiency unless (5) holds. See Appendix.)

If the cell frequencies are merely *very roughly* proportionate the reasoning described above for the proportionate case will apply approximately, and suitable unequal weights may be expected to be more efficient than equal weights, and to be nearly as efficient as using the additive model if the within-cell *df* are not too small. (With the sample sizes of Example 2, for instance, weights proportional to the marginal cell frequencies are 91% as

efficient as the optimal weights.)

Of course, in practice many other considerations besides efficiency may influence the choice of weights; for instance, if several studies of the treatment factor A are to be compared, common weights should be used for all the studies.

5. IMPLEMENTATION OF WEIGHTED ANOVA

The full-model F-test for H_A with arbitrary weights $\{w_j\}$ and arbitrary cell frequencies $\{n_{ij}\}$ is straightforward to compute. The numerator SS for the F-statistic (3) is (Scheffé 1959, page 118):

$$\begin{aligned} SSA &= \sum_i U_i(\hat{A}_i - \overline{\hat{A}}_i)^2 \\ &= \Sigma U_i \hat{A}_i^2 - (\Sigma U_i \hat{A}_i)^2/(\Sigma U_i), \end{aligned} \tag{8}$$

where

$$U_i = (\sum_j w_j^2/n_{ij})^{-1}$$

and

$$\overline{\hat{A}} = (\Sigma U_i \hat{A}_i)/(\Sigma U_i).$$

(In the case of proportionate cell frequencies, the value of SSA using the optimal weights (6) is the same as would be computed for a one-way ANOVA on Factor A.) If $L = \sum c_i A_i$ is any contrast in the $\{A_i\}$, the least-squares estimate of L in the full model is $\hat{L} = \sum c_i \hat{A}_i$ and the squared standard error of $\hat{L}$ is

$$s_{\hat{L}}^2 = (MSW) \sum_i \sum_j (c_i^2 w_{ij}^2/n_{ij}).$$

Among programs available in statistical computing packages, BMDP4V is noteworthy for the ease with which the user can specify weights $\{w_j\}$. In SAS GLM an analysis with specified $\{w_j\}$ can be requested with the CONTRAST statement. SPSS^X does not permit the use of unequal weights.

6. EXTENSIONS

In addition to Factors A and B, an investigation may include other factors and/or covariates $Z_1, Z_2, \ldots,$. If it is assumed that there are no interactions between the $\{Z_i\}$ and Factors A and B, then the discussion in Sections 3 and 4 extends directly. If such interactions are contemplated in the model, then the problem of choosing weights may be considerably more complicated, although analogous ideas can be applied.

7. CONCLUSION

In the two-way ANOVA with Factor B a nuisance factor, analysis with unequal weights is conceptually simple, is computationally easy, and can greatly increase efficiency; it should receive more emphasis in the training of applied statisticians and should be included as a convenient option in computing packages.

APPENDIX

Proof of Proposition 1

It is well-known (Scheffé) 1959, pp. 38–39 and p. 118) that the statistic (3) has a noncentral F-distribution with non-centrality parameter δ^2 satisfying

$$\sigma^2\delta^2 = \sum_i U_i(A_i - \overline{A})^2, \tag{9}$$

where $\overline{A} = (\Sigma U_i A_i)/(\Sigma U_i)$. If H_{AB} is true, then μ_{ij} can be written as

$$\mu_{ij} = \mu + \alpha_i^* + \beta_j^*, \tag{10}$$

where $\Sigma\alpha_i^* = 0$ and $\Sigma\beta_j^* = 0$; note that $\{\alpha_i^*\}$ and $\{\beta_j^*\}$ are defined independently of any weights. Using (5) and (10), (9) reduces to

$$\sigma^2\delta^2 = (Gn_{++})^{-1}\sum_i n_{i+}(\alpha_i^* - \overline{\alpha}^*)^2, \tag{11}$$

where $\overline{\alpha}^* = n_{++}^{-1}\Sigma n_{i+}\alpha_i^*$, and

$$G = \sum_j (w_j^2/r_j).$$

The quantity (11) depends on $\{w_j\}$ only through G, and it is straightforward to show that G is minimized for the weights (6). Similarly, when (5) holds, the sampling variance of any estimated contrast is proportional to G. The relative efficiency (7) is the ratio of the corresponding values of G.

□

Proof of Proposition 2

Define the vector $\mathbf{\Psi} = [(\alpha_2 - \alpha_1), (\alpha_3 - \alpha_1), \ldots, (\alpha_a - \alpha_1)]'$, where the $\{\alpha_i\}$ are defined using weights $\{w_j\}$ given by (6) and any weights $\{v_i\}$. The SS for testing H_A in the full model can be written (Scheffé 1959, p. 40):

$$SSA = \hat{\mathbf{\Psi}}'\mathbf{B}^{-1}\hat{\mathbf{\Psi}}$$

where $\hat{\mathbf{\Psi}}$ is the least-squares estimate of $\mathbf{\Psi}$ in the full model, and $\mathbf{B} = \sigma^{-2}\ \mathrm{Var}(\hat{\mathbf{\Psi}})$. The SS for testing H_A in the additive model is

$$SSA^* = \hat{\mathbf{\Psi}}^{*\prime}\mathbf{B}^{*-1}\hat{\mathbf{\Psi}}^*$$

where $\hat{\mathbf{\Psi}}^*$ is the least-squares estimate of $\mathbf{\Psi}$ in the additive model, and $\mathbf{B}^* = \sigma^{-2}\ \mathrm{Var}(\hat{\mathbf{\Psi}}^*)$. (The value of $\hat{\mathbf{\Psi}}^*$ is independent of the weights.) It is easy to show (Scheffé 1959, pp. 117–119) that if (5) holds then $\hat{\mathbf{\Psi}} = \hat{\mathbf{\Psi}}^*$, which proves Proposition 2.

□

Non-proportionate Cell Frequencies

If $a = 2$ then it follows easily from (9) and (11) that δ^2 is maximized for weights $\{w_j\}$ proportional to the *harmonic* means of the cell frequencies; further, it is straightforward to show that the analog of Proposition 2 is true for these weights. Nevertheless, these weights are not desirable because the corresponding A_i are not readily interpretible if H_{AB} is false. If $a \geq 3$, the weights $\{w_j\}$ which maximize (9) will in general depend on the parameter point $(\alpha_1^*, \ldots, \alpha_a^*)$.

REFERENCES

SCHEFFÉ, H. (1959), *The Analysis of Variance*, New York: Wiley.

SEARLE, S.R., SPEED, F.M., AND MILLIKEN, G.A. (1980), "Population marginal Means in the Linear Model: An Alternative to Least Squares Means," *The American Statistician*, 34, 216–221.

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.

STEINHORST, R.K. (1982), "Resolving Current Controversies in Analysis of Variance," *The American Statistician*, 36, 138–139.