On Presence of Interaction In An Unbalanced Two-Way Random Model

[1,] F.C. Eze, [2,] P.E. Chigbu
[1.] Department of Statistics, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria
[2.] Department of Statistics, University of Nigeria, Nsukka, Nigeria

Abstract

In an unbalanced two-way random model, there is no obvious denominator for testing for the main effects as a result of the presence of inter-action. The interaction was removed from the model/data resulting to a reduced model devoid of interaction. The data were transformed by dividing the entries of each cell of the data by the inverse of the square root of the standard error to remove the interaction.

Keywords: Expected mean squares, fractional degrees of freedom, reduced model.

1. Introduction

The presence of interaction in two-way analysis of variance could be a serious problem when testing for the main effects. Chow[3] emphasized that testing the treatment effects for the two-way ANOVA model with interaction, the interaction effect has to be tested first, otherwise the result for testing the treatment cannot be interpreted in a statistical meaningful manner. In ANOVA, a large F -value provides evidence against the null hypothesis. However, the interaction test should be examined first. The reason for this is that, there is little point in testing the null hypothesis H_A or H_B if H_{AB}: no interaction effect is rejected, since the difference between any two levels of a main effect also includes an average interaction effect Cabrera and McDougall[1] argued. Moore et al [5] argued that, there are three hypotheses in a two-way ANOVA, with an F -test for each. We can test for significance of the main effect A, the main effects of B and AB interaction. It is generally a good practice to examine the test interaction first, since the presence of a strong interaction may influence the interpretation of the main effects. Muller and Fetterman [6] agreed that model reduction in the presence of non significant interaction may be attractive for unbalanced or incomplete design.

2. Methodology

Given the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + e_{ijk} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n_{ij} \end{cases} \qquad (1)$$

Where

y_{ijk} is the kth observation in ijth cell,

μ is the overall mean effect,

λ_{ij} is the effect of the interaction between factor A and factor B,

e_{ijk} is a random error components,

n_{ij} is the number of observation per cell; and

using the Brute-Force Method, the expected mean squares (EMS) of the parameters of Equation (1) can be shown to be

$$EMS_A = \sigma_e^2 + k_\alpha \sigma_\alpha^2 + k_1 \sigma_\lambda^2$$
$$EMS_B = \sigma_e^2 + k_\beta \sigma_\beta^2 + k_2 \sigma_\lambda^2$$
$$EMS_\lambda = \sigma_e^2 + k_3 \sigma_\lambda^2$$
$$EMS_e = \sigma_e^2$$

where

$$k_\alpha = \frac{N - N^{-1} \sum_i N_i^2}{a-1} \qquad (2)$$

$$k_1 = \frac{(\sum_i N_i^{-1} \sum_j n_{ij}^2 - N^{-1} \sum_{ij} n_{ij}^2}{a-1} \qquad (3)$$

$$k_\beta = \frac{N - N^{-1} \sum_j N_j^2}{b-1} \qquad (4)$$

$$k_2 = \frac{(\sum_j N_j^{-1} \sum_i n_{ij}^2 - N^{-1} \sum_{ij} n_{ij}^2}{b-1} \qquad (5)$$

$$k_3 = (N - \sum_i N_i^{-1} \sum_j n_{ij}^2 - \sum_j N_j^{-1} \sum_i n_{ij}^2 - N^{-1} \sum_{ij} n_{ij}^2$$
$$+ 2\sum_{ij} n_{ij}^3 N_i^{-1} N_j^{-1}) / (a-1)(b-1) \qquad (6)$$

where

$$\sigma_\alpha^2 = E(\bar{y}_{i..} - \bar{y}_{..})^2 = \frac{\sum_i (\bar{y}_{i..} - \bar{y}_{...})^2}{n}$$

$$\sigma_\beta^2 = E(\bar{y}_{.j.} - \bar{y}_{..})^2 = \frac{\sum_i (\bar{y}_{.j.} - \bar{y}_{...})^2}{n}$$

$$\sigma_\lambda^2 = E(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 = \frac{\sum_{ij} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2}{n}$$

$$\sigma_e^2 = E(y_{ijk} - \bar{y}_{ij.})^2 = \frac{\sum_{ijk} (y_{ijk} - \bar{y}_{ij.})^2}{n}$$

k_1, k_2 and k_3 are the coefficient of the variance components of the interaction for factor A, factor B and the interaction between factor A and factor B respectively.

The expected mean squares (EMS) of Equation (1) are presented in the ANOVA Table shown in Table 1.

S.V	d.f	SS	MS	EMS
Factor A	a-1	SS_α	MS_α	$\sigma_e^2 + k_\alpha \sigma_\alpha^2 + k_1 \sigma_\lambda^2$
Factor B	b-1	SS_β	MS_β	$\sigma_e^2 + k_\beta \sigma_\beta^2 + k_2 \sigma_\lambda^2$
AxB	(a-1)(b-1)	SS_λ	MS_λ	$\sigma_e^2 + k_3 \sigma_\lambda^2$
Error	N-pq	SS_e	MS_e	σ_e^2
Total	N-1	SS_T		

Table 1: ANOVA Table for Unbalanced data.

From the expected mean squares in Table 1, we can see that the appropriate statistic for testing the no interaction hypothesis $H_0 : \sigma_\lambda^2 = 0$ is

$$F_c = \frac{MS_\lambda}{MS_e}$$

This is because, under H_0 both numerator and denominator of F_c have expectation σ_e^2.

The case is different when testing for $H_0 : \sigma_\alpha^2 = 0$ because the numerator expectation is $\sigma_e^2 + k_1 \sigma_\lambda^2$ and no other expectation in Table 1 that is $\sigma_e^2 + k_1\sigma_\lambda^2$ under H_0.

The case is also the same when testing for $H_0 : \sigma_\beta^2 = 0$

The cause of the above problem is the presence of interaction and when the interaction is removed from the model, the ANOVA Table 1 reduces to the ANOVA Table shown in Table 2.

S.V	d.f	SS	MS	EMS
Factor A	a-1	SS_α	MS_α	$\sigma_e^2 + k_\alpha\sigma_\alpha^2 + k_1\sigma_\lambda^2$
Factor B	b-1	SS_β	MS_β	$\sigma_e^2 + k_\beta\sigma_\beta^2 + k_2\sigma_\lambda^2$
Error	N-a-b+1	SS_e	MS_e	σ_e^2
Total	N-1	SS_T		

Table 2: Reduced ANOVA Table for Unbalanced data.

From the ANOVA Table 2, it is obvious that the common denominator for testing for the main effects is MS_e.

3 Method of Removing the Interaction

Now to remove the interaction from the model we proceed as follows:

The least square estimate for the interaction is:

$$\lambda_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

$\bar{y}_{ij.}, \bar{y}_{i..}, \bar{y}_{.j.}, \bar{y}_{...}$ are the cell means, row means for factor A, column means for factor B and the overall means respectively for the full model.

Removing the interaction from Equation (1) we have

$$y_{ijk}^* = y_{ijk} - \bar{y}_{ij.} + \bar{y}_{i..} + \bar{y}_{.j.} - \bar{y}_{...} \qquad (7)$$

$$= \mu + \alpha_i + \beta_j + z_{ijk}$$

Dividing Equation (7) by k and simplifying we have

$$y_{ij.}^* = \bar{y}_{i..} + \bar{y}_{.j.} - \bar{y}_{...}$$

Similarly

$$y_{i..}^* = \bar{y}_{i..}$$

and

$$y_{...}^* = \bar{y}_{...}$$

z_{ijk} is the error associated with y_{ijk}^* for the reduced model while $\bar{y}_{ij.}^*, \bar{y}_{i..}^*$ and $\bar{y}_{...}^*$ are the cell means, row means for factor A and the overall means for the reduced model.

It then follows that

$$\bar{y}_{i..}^* - \bar{y}_{...}^* = \bar{y}_{i..} - \bar{y}_{...}$$
$$\bar{y}_{.j.}^* - \bar{y}_{...}^* = \bar{y}_{.j.} - \bar{y}_{...}$$

Now

$$z_{ijk} = e_{ijk} - \bar{e}_{ij.} + \bar{e}_{i..} + \bar{e}_{.j.} - \bar{e}_{...}$$

To transform and normalize the data, we divide the original data by the inverse of the standard error. To do this we need the $\operatorname{var}(z_{ijk})$.

From Equation (1) n_{ij} is the number of observations per cell. It then follows that

$$\sum_j n_{ij} = N_{i.}$$

$$\sum_i n_{ij} = N_{.j}$$

$$\sum_{ij} n_{ij} = N_{..}$$

It can be shown that

$$E\left(\bar{e}_{i..}\right) = \frac{\sigma_e^2}{N_{i.}}$$

$$E\left(\bar{e}_{.j.}\right) = \frac{\sigma_e^2}{N_{.j}}$$

$$E\left(\bar{e}_{ij.}\right) = \frac{\sigma_e^2}{n_{ij}}$$

$$E\left(\bar{e}_{...}\right) = \frac{\sigma_e^2}{N_{..}}$$

Therefore

$$\text{Variance of } z_{ijk} = \left(\sigma_e^2 - \frac{\sigma_e^2}{n_{ij}} + \frac{\sigma_e^2}{N_{i.}} + \frac{\sigma_e^2}{N_{.j}} + \frac{2n_{ij}\sigma_e^2}{N_{i.}N_{.j}} - \frac{3\sigma_e^2}{N_{..}}\right)$$

$$= \sigma_e^2\left(1 - \frac{1}{n_{ij}} + \frac{1}{N_{i.}} + \frac{1}{N_{.j}} + \frac{2n_{ij}}{N_{i.}N_{.j}} - \frac{3}{N_{..}}\right)$$

Therefore

$$E\left(MS_z\right) = k_{ij}E\left(MS_e\right) = k_{ij}\sigma_e^2$$

Where

$$k_{ij} = \left(1 - \frac{1}{n_{ij}} + \frac{1}{N_{i.}} + \frac{1}{N_{.j}} + \frac{2n_{ij}}{N_{i.}N_{.j}} - \frac{3}{N_{..}}\right) \qquad (8)$$

Therefore

$$SS_z = \sum_{ijk} \frac{1}{\sqrt{k_{ij}}}\left(y_{ijk} - \bar{y}_{ij.}\right)^2 = \sum\left(w_{jk} - \bar{w}_{ij.}\right)^2$$

Where

$$w_{ijk} = \frac{y_{ijk}}{\sqrt{k_{ij}}}$$

w_{ijk} is now the transformed data.

4 Illustrative Example

The data below correspond to an experiment in which four different methods for growing crops were tested on four different types of fields (same soil but different light exposure). The yield was measured after the harvest. Because the 3rd method was not tested on the 4th type of field (because of a lack of seeds), and the 2nd method on the 4th type of field (because of a hail storm), the experiment is a typical example of an unbalanced ANOVA.

		Type of field		
Method	1	2	3	4
1	20, 7	39, 17	34, 13	13, 5
2	35, 52	30, 28	58, 73	64
3	62, 44	82, 81	69, 84	-

Table 3: Source: Kovach Computing Services [4]

The model is the same as in Equation (1), where

y_{ijk} the yield after harvest.

μ is a constant.

α_i is the mean effects of the method representing factor A.

β_j is the mean effects of the type of field representing factor B.

λ_{ij} is the interaction between the method and type of field.

e_{ijk} is the error associated with y_{ijk} .

The sums of squares for the unbalanced data which are analogous to balanced designs are:

$$SS_A = \sum_i N_i \left(\bar{y}_{i..} - \bar{y}_{...}\right)^2 \quad (9)$$

$$SS_B = \sum_j N_j \left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2 \quad (10)$$

$$SS_\lambda = \sum_{ij} n_{ij} \left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}\right)^2 \quad (11)$$

$$SS_e = \sum_{ijk} (y_{ijk} - \bar{y}_{ij.})^2 \quad (12)$$

Using Equations (9), (10), (11) and (12), the sums of squares divided by their respective degrees of freedom for factor A, factor B, the interaction between factor A and factor B and the error term are respectively:-
MS_A = 4749.22, MS_B = 641.21, MS =458.7 and MS_e = 123.
Testing for the main effects will need a special F -test because of the unbalanced nature of the design. However, if interaction is non significant or absent, the common denominator for testing for the main effects is the mean square error. Testing for the interaction we have:

$$F = \frac{MS_\lambda}{MS_e} = \frac{458.7}{123} = 3.73$$

From $F_{6,9}^{0.05}$ we have 3.37 and since 3.73 > 3.37, this shows the presence of interaction in the data, hence the need for appropriate transformation of data.

Using equation 8 the transformed data are shown in Table 4.

	Type of field			
Method	1	2	3	4
1	17.2, 6.02	33.5, 14.6	29.2, 11.2	9.1, 3.5
2	30.5, 45.2	26.1, 24.4	50.5, 63.5	41.6
3	55.2, 39.2	73, 72.1	61.4, 74.8	-

Table 4: Transformed unbalanced data

Having transformed the data, the new sums of squares for the main effects and error are:

$$SS_A = \sum_i N_i \left(\bar{y}_{i..} - \bar{y}_{...}\right)^2 \quad (13)$$

$$SS_B = \sum_j N_j \left(\bar{y}_{.j.} - \bar{y}_{...} \right)^2 \tag{14}$$

$$SS_e = \sum_{ijk} (y_{ijk} - \bar{y}_{ij.})^2 \tag{15}$$

The sums of squares and the corresponding mean sum of squares for the main effects and the error terms has been calculated using Equations (13), (14) and (15) and presented in the ANOVA Table 5

S.V	d.f	SS	MS	F -ratio
Method	2	7705.35	3852.68	69.54
Type of field	3	2066.43	688.81	12.43
Error	15	831.01	92.33	
Total	20	28602.79		

Table 5: ANOVA Table

From $F_{2,15}^{0.05}$ = 3:68 and $F_{3,15}^{0.05}$ = 3:29, the main effects are significant.

5 Summary and Conclusion

We have shown that in an unbalanced two-way random model, there are no obvious denominator for testing for the main effects which is as a result of the presence of interactions.

When interaction is present in an unbalanced two-way random model, we need to construct an appropriate F -test for the main effects. To avoid such situation, the interaction from the data/model should be removed to have a valid result when testing for the main effects.

References

[1] [Cabrera and McDougall 2002] Cabrera, J. and McDougall, A. (2002). Statistical Consulting. Springer.

[2] [Castrup 2010] Castrup, H. (2010). A Welch-Satterthwaite relation for correlated errors.Integrated sciences crop Bakers eld CA 93306.

[3] [2003] Chow, S.C. (2003) Encyclopedia of Biopharmaceutical Statistics. p 1027. Informa Health Care.

[4] [2011] Kovach Computing Services (2011). "How can I use XLSTAT to run a two-way unbalanced ANOVA with interaction." Anglesey, Wales.

[5] [2004] Moore, D.S., McCablee, G., Duckworth, W., and Sclove, S. (2004). Practice of business Statistics, part iv pp 15-17. Palgrave Macmillan.

[6] [Muller and Fetterman 2002] Muller, K., and Bethel, F. (2002). Regression and Analysis of Variance. An integrated approach using SAS P.381. SAS publishing.