Journal of Natural Sciences Research
ISSN 2224-3186 (Paper) ISSN 2225-0921 (Online)
Vol.2, No.10, 2012

Unbalanced Two-Way Random Model with Integer-Valued Degrees of Freedom

F.C. Eze[1*] P.E. Chigbu[2]

1. Department of Statistics, Nnamdi-Azikiwe University, Awka, Nigeria
2. Department of Statistics University of Nigeria, Nsukka, Nigeria

[*]E-mail: ezefcc@yahoo.com

Abstract

The expected mean squares for unbalanced two-way random model were de-rived. From the ANOVA table, $k_1 \neq k_2$ and as such there is no obvious denominator for testing for the main effects. k_1 and k_2 are $\frac{(\sum_i N_i^{-1}\sum_j n_{ij}^2 - N^{-1}\sum_{ij} n_{ij}^2}{a-1}$ and $\frac{(\sum_j N_j^{-1}\sum_i n_{ij}^2 - N^{-1}\sum_{ij} n_{ij}^2}{b-1}$ which are the coefficients of the variance components of the interaction for factor A and factor B respectively. A theorem was proved to show that if $k_1 = k_2$, the unbalanced data becomes balanced and the common denominator for testing for the main effects becomes the mean square error.

Keywords: Expected mean squares, fractional degrees of freedom

1. Introduction

Unbalanced data in two-way layout are such that the numbers of observations in each cell of the layout are not the same which include cases where there are no observations in some cell. Analyzing the variance of data which are classified in two-ways with unequal numbers of observations falling into each cell of the classification needs special methods of analysis because of the inequality of the cell numbers.

Malwane and Samaradasa (1997) considered two-way ANOVA model with unequal cell frequencies without the assumption of equal error variances by taking generalized approach to finding p-values. The generalized F -test they developed in their article can be utilized in significance testing or in fixed level testing under the Neyman-Pearson theory which they claim has its advantage over the classical F -test.

Montgomery (2001) gave an easy case of analyzing unbalanced data when the classifications of the data are proportional. In this case, the number of

observations in the ijth cell is

$$n_{ij} = \frac{N_{i.}N_{.j}}{N_{..}}$$

This condition implies that the numbers of observations in any two rows or columns are proportional. When this situation occurs, the standard analysis of variance can be employed with minor modification in the computing formulas for the sums of squares.
Montgomery (2001) however did not indicate if the significant testing involves fixed, random or mixed effect testing.

The two-way crossed classification with interaction for mixed model when data are unbalanced is a problem especially when deriving the expected mean squares and the variance components. When the expected mean squares for the mixed model are derived, it always contains the functions of the fixed effects. The functions of the fixed effects cannot be eliminated by considering linear combinations of the mean squares and expected mean squares. It then becomes practically impossible to determine the appropriate F -ratio for testing for the main effects.

In our paper, the expected mean squares for the random model for such lay-out were derived and appropriate F -ratio determined. The number of degrees of freedom for the denominator of the derived F -ratio is non-integer-valued. Approximating the non-integer degrees of freedom creates an uncertainty regarding the exact value of the F -ratio from the F -distribution table. If $k_1 = k_2$, the unbalanced data becomes balanced and the main effects can be tested using the mean square error as the common denominator for the F -ratio.

2. Methodology

Given the model

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

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

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

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

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

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

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

where

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

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

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

3

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

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_A	MS_A	$\sigma_e^2 + k_{\alpha}\sigma_{\alpha}^2 + k_1\sigma_{\lambda}^2$
Factor B	b-1	SS_B	MS_B	$\sigma_e^2 + k_{\beta}\sigma_{\beta}^2 + k_2\sigma_{\lambda}^2$
AxB	(a-1)(b-1)	SS_{λ}	MS_{λ}	$\sigma_e^2 + k_3\sigma_{\lambda}^2$
Error	N-pq	SS_e	SS_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 $H_0 : \sigma_{\beta}^2 = 0$

From Table 1,if we were interested in testing for the main effects A and B using F -test Statistic, there would be no obvious denominator for testing the hypotheses $H_0 : \sigma_{\alpha}^2 = 0$ and $H_0 : \sigma_{\beta}^2 = 0$ because $k_1 \neq k_2$.

If we are interested in testing $H_0 : \sigma_{\alpha}^2 = 0$ and we can find a linear combination of independent mean squares with expectation:

$$MS\phi_1 = \sigma_e^2 + k_1\sigma_{\lambda}^2 \qquad (7)$$

we would have F -test for the null hypothesis given by

$$F_{f_{\alpha}, f_{\phi}} = \frac{MS_{\alpha}}{MS\phi_1}$$

f_{α} and f_{ϕ} are the degrees of freedom for the numerator and denominator of the F -ratio respectively.

$MS\phi_1$ is obtained from Table 1 as follows:

$$\sigma_e^2 = MS_e$$

$$MS_\lambda = \sigma_e^2 + k_3\sigma_\lambda^2$$

Thus

$$MS_\lambda = MS_e + k_3\sigma_\lambda^2$$

This implies

$$\sigma_\lambda^2 = \frac{MS_\lambda - MS_e}{k_3} \tag{8}$$

Substituting Equation (8) in Equation (7) gives

$$MS\phi_1 = \left[MS_e + k_1\left(\frac{MS_\lambda - MS_e}{k_3}\right)\right]$$

$$= MS_e + \left(\frac{k_1}{k_3}\right)MS_\lambda - \left(\frac{k_1}{k_3}\right)MS_e$$

Therefore

$$MS\phi_1 = (1-\theta_1)MS_e + (\theta_1)MS_\lambda \tag{9}$$

where

$$\theta_1 = \frac{k_1}{k_3}$$

From Castrup (2010) the degree of freedom for the denominator is:

$$f_{\phi_1} = \frac{(MS\phi_1)^2}{(1-\theta_1)^2\frac{MS_e^2}{f_e} + (\theta_1)^2\frac{MS_\lambda}{f_\lambda}} \tag{10}$$

Similarly, testing $H_0 : \sigma_\beta^2 = 0$

$$MS\phi_1 = (1-\theta_2)MS_e + (\theta_2)MS_\lambda$$

Similarly from Castrup (2010), the degree of freedom for the denominator is

$$f_{\phi_2} = \frac{(MS\phi_2)^2}{(1-\theta_2)^2\frac{MS_e^2}{f_e} + (\theta_2)^2\frac{MS_\lambda}{f_\lambda}} \tag{11}$$

$$\theta_2 = \frac{k_2}{k_3}$$

where

f_e and f_λ are the degree of freedom obtained from the ANOVA Table 1

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

$$SS_A = \sum_i N_i\left(\bar{y}_{i..} - \bar{y}_{...}\right)^2 \tag{12}$$

$$SS_B = \sum_j N_j\left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2 \tag{13}$$

$$SS_{\lambda} = \sum_{ij} n_{ij} \left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...} \right)^2 \quad (14)$$

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

Equations (10) and (11) are non-integer-valued degrees of freedom. This means we need to correct to the nearest whole number before reading the value from our F -distribution table. However, if $k_1 = k_2$, the problem of fractional degrees of freedom will be solved.

Theorem:

Given $H_0 : \sigma_{\alpha}^2 = 0,$ and $H_0 : \sigma_{\beta}^2 = 0,$ then $F = \dfrac{MS_{\alpha}}{MS_e} = \dfrac{MS_{\beta}}{MS_e}$ if and only if $k_1 = k_2$.

Proof:

From Table 1

$$MS_{\alpha} = \sigma_e^2 + k_{\alpha}\sigma_{\alpha}^2 + k_1\sigma_{\lambda}^2$$
$$MS_{\beta} = \sigma_e^2 + k_{\beta}\sigma_{\beta}^2 + k_2\sigma_{\lambda}^2$$

and

$$MS_e = \sigma_e^2$$

Thus

$$MS_{\alpha} = \sigma_e^2 + k_{\alpha}\sigma_{\alpha}^2 + k_1\sigma_{\lambda}^2$$

and

$$MS_{\beta} = \sigma_e^2 + k_{\beta}\sigma_{\beta}^2 + k_2\sigma_{\lambda}^2$$

This implies that

$$\frac{MS_{\alpha}}{MS_e} = 1 + \frac{k_{\alpha}\sigma_{\alpha}^2 + k_1\sigma_{\lambda}^2}{MS_e} \text{ and } \frac{MS_{\beta}}{MS_e} = 1 + \frac{k_{\beta}\sigma_{\beta}^2 + k_2\sigma_{\lambda}^2}{MS_e} \quad (16)$$

Now, suppose $k_1 = k_2$, then we obtain from Equation (16)

$$\frac{MS_{\alpha}}{MS_e} - \frac{MS_{\beta}}{MS_e} = \frac{k_{\alpha}\sigma_{\alpha}^2 - k_{\beta}\sigma_{\beta}^2}{MS_e} \quad (17)$$

Thus, testing that $\sigma_{\alpha}^2 = \sigma_{\beta}^2 = 0$, we obtain from Equation (17), that

$$\frac{MS_{\alpha}}{MS_e} = \frac{MS_{\beta}}{MS_e}$$

Conversely, suppose

$$\frac{MS_\alpha}{MS_e} = \frac{MS_\beta}{MS_e}$$

That is, suppose

$$\sigma_e^2 + k_\alpha \sigma_\alpha^2 + k_1 \sigma_\lambda^2 = \sigma_e^2 + k_\beta \sigma_\beta^2 + k_2 \sigma_\lambda^2$$

then, testing that $\sigma_\alpha^2 = \sigma_\beta^2 = 0$ gives

$$k_1 \sigma_\lambda^2 = k_2 \sigma_\lambda^2$$

This implies that $(k_1 - k_2)\sigma_\lambda^2 = 0$

but since $\sigma_\lambda^2 \neq 0$, then $k_1 - k_2 = 0$, that is $k_1 = k_2$.

This completes the proof.

3. Estimation of missing values

There are several methods of estimating missing values in unbalanced data. Howell (2008) gave some methods of treating missing data. Some of the methods are Mean substitution, Regression substitution, Listwise deletion and Pair-wise deletion.

Little and Rubin (2002) gave some classifications of missingness. We have

(a) Missingness at random (MCAR): This is a situation where the probability of misssing data does not depend on observed or unobserved data.

(b) Missing at random(MAR): Here the probability of missing data does not depend on the unobserved data but conditional on the observed data.

(c) Missing not at random(MNAR):Here the probability of missing data does depend on the unobserved, conditional on the observed data.

Irrespective of the classifications, our interest is to estimate missing value(s) to set $k_1 = k_2$ using any appropriate method of estimation. Each estimation will lead to subtracting one from the error mean square. See Montgomery (2001).

4. Illustrative example

Suppose an oil company gets its crude oil from 3 different sources and refines it in 3 different refineries. In one part of the refining process, a measurement of efficiency is taken as a percentage and recorded as an integer between 0 and 100. The data are shown below:

Refinery	Texas	Oklahoma	Gulf of Mexico
Galveston	31	36,38	26
Newark	39,59	37,36	42
Savanna	42,44	36,42	26,37

Table 2: Crude oil data from Searle (1997), p.162

The model is of the type in Equation (1), where,

y_{ijk} is the measurement of the efficiency,

μ is the overall mean,

α_i is the average effects of the refineries,

β_j is the average effects of the sources of the crude oil,

λ_{ij} is the interaction between the refineries and the sources of the crude oil and

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

From Equations (3) and (5)

$$k_1 = k_2 = 1.75$$

and from Equation (6)

$$k_3 = 1.6$$

Using Equations (12), (13), (14) and (15), the mean sums of squares for factor A, factor B, the interaction between factor A and B and the error term has been calculated and presented below.

$MS_\alpha = 108.08;\ MS_\beta = 118.34;\ MS_\lambda = 31.52$ and $MS_e = 47.17.$

Since $k_1 = k_2,$ and $\theta_1 = \theta_2 = 1.09$

This implies that

$$MS\phi_1 = MS\phi_2$$

From the above Theorem, our hypotheses shall be:

$H_{01} : \sigma_\alpha^2 = 0$ versus the alternative, H_{11} : at least one of the variances differ.

Thus

$$F = \frac{MS_\alpha}{MS_e} = \frac{108.08}{47.17} = 2.29$$

From the F-distribution table, we have $F_{2,6}^{0.05} = 5.14$. Since 2.29 < 5.14 the variances are non-significant.

Similarly, $H_{02} : \sigma_\beta^2 = 0$ versus the alternative, H_{12} : at least one of the variances differ.

$$F = \frac{MS_\beta}{MS_e} = \frac{118.34}{47.17} = 2.51$$

From the F-distribution table, we have $F_{2,6}^{0.05} = 5.14$ and since 2.51 < 5.14 the variances are also non-significant.

Finally, for $H_{03} : \sigma_\lambda^2 = 0$ versus the alternative, H_{13} : at least one differs,

$$F = \frac{MS_\lambda}{MS_e} = \frac{31.52}{47.17} = 0.67$$

From the F-distribution table, we have $F_{4,6}^{0.05} = 4.53$, and since 0.67 < 4.53 the variances are non-significant.

5. Summary and Conclusion

We have seen that when $k_1 \neq k_2$, there are no obvious denominator for testing for the main effects in an unbalanced two-way random model. This is as a result of the presence of interaction. From our theorem, when $k_1 = k_2$, the interaction is removed from the data and the main effects can then be tested using MS_e as the denominator of our F -test which has an integer-valued degree of freedom.

In a situation where $k_1 \neq k_2$, the missing values should be estimated to set $k_1 = k_2$ so as to avoid testing for the main effects when interaction is present.

For valid results, it is generally advisable to test for the main effects in an unbalanced two-way random model when interaction is absent.

Acknowledgment

The authors wish to thank Mrs. Happiness Ilouno Obiorah of Department of Statistics, Nnamdi-Azikwe University, Awka, Nigeria for introducing us to the Journal of Natural Sciences Research and her valuable contributions.

References

Castrup, Howard (2010), "A Welch-Satterthwaite relation for correlated errors", Integrated sciences crop Bakersfield CA 93306.

Howell, D.C. (2008), The analysis of missing data , In Outhwaite, W. and Turner, S. Handbook of Social Science Methodology, London. Sage.

Little, R.J.A and Rubin, D.B. (2002), Statistical Analysis with missing Data , Second edition. Hoboken: Wiley.

Malwane, M.A. and Samaradasa, W. (1997), "Two-way ANOVA with unequal cell frequencies and unequal variances." Statistica Sinica 7, 631-646.

Molenbergehs, G., Beunckens, C. and Sotto, C. (2008), " Every missingness not at random model has a missingness at random counterpart with equal fit" , J. R. Statist. Soc. B, 70, pp 371-388.

Montgomery, D.C. (2001), Design and Analysis of Experiment ,John Willey and Sons.

Searle, S.R. (1997). Linear models, John Willey and sons, Inc

This academic article was published by The International Institute for Science, Technology and Education (IISTE). The IISTE is a pioneer in the Open Access Publishing service based in the U.S. and Europe. The aim of the institute is Accelerating Global Knowledge Sharing.

More information about the publisher can be found in the IISTE's homepage: http://www.iiste.org

CALL FOR PAPERS

The IISTE is currently hosting more than 30 peer-reviewed academic journals and collaborating with academic institutions around the world. There's no deadline for submission. **Prospective authors of IISTE journals can find the submission instruction on the following page:** http://www.iiste.org/Journals/

The IISTE editorial team promises to the review and publish all the qualified submissions in a **fast** manner. All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Printed version of the journals is also available upon request from readers and authors.

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar