Science Journal of Applied Mathematics and Statistics
2015; 3(6): 288-292
Published online December 25, 2015 (http://www.sciencepublishinggroup.com/j/sjams)
doi: 10.11648/j.sjams.20150306.18
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

Two Factor Data Analysis with Unequal Cell Frequencies and Interaction

Chinwendu Alice Uzuke*, Ikewelugo Cyprian Anene Oyeka, Happiness Onyebuchi Obiora-Ilouno

Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria

Email address:
uzuke.ca@gmail.com (C. A. Uzuke)

To cite this article:
Chinwendu Alice Uzuke, Ikewelugo Cyprian Anene Oyeka, Happiness Onyebuchi Obiora-Ilouno. Two Factor Data Analysis with Unequal Cell Frequencies and Interaction. *Science Journal of Applied Mathematics and Statistics.* Vol. 3, No. 6, 2015, pp. 288-292.
doi: 10.11648/j.sjams.20150306.18

Abstract: This paper proposes a non parametric method for two factor data analysis with unequal cell frequencies and interaction. Chi-square test statistic was developed for testing the null hypothesis of no treatment effect and interaction between factor A and factor B. The proposed methods are illustrated with some data and compared with the usual unweighted mean method. The result showed that the proposed method is more powerful than the method of unweighted mean.

Keywords: Cell Frequency, Interaction, Chi-square, Unweighted Mean, Ranking, Tied Observation

1. Introduction

Analysis of variance (ANOVA) is generally regarded as the best analysis technique for balanced experiments that have equal number of subjects in each group that is cells with equal frequency [3]. Just as it may often be too difficult and too expensive to obtain more than one observation per treatment combination, it may also prove impossible to obtain equal number of observation per cell in a two factor analysis. For example, even though an experiment was planned with equal number of observations per cell. Some of the observation may end up missing for various reasons.

Some classifications of missingness were given as missingness at random (MCAR) as a situation where the probability of missing data does not depend on observed or un observed data, missing at random (MAR) as the probability that the missing data does not depend on the observed data while missing not at random (MNAR), is the probability that the missing data depends on the unobserved data conditional on the observed data [4]. Data with unequal cell frequency are not too far from those with equal frequency, it is sometimes possible to use approximate procedures that convert the former from the later. In practice the decision must be made when data are not sufficiently different from the case with equal frequency which makes the degree of approximation introduced relatively unimportant. [1]

Two – way ANOVA with unequal cell frequencies without assumption of equal error variance was considered by taking generalized approach to finding p-values [5]. But when the sample size per treatment combination is not the same for all treatments in a two factor ANOVA, the factor effect become more complicated and the usual calculations are no longer directly applicable [7], [9]. In this situation, the easiest and exact way to obtain the proper sum of squares for testing factor effects and interactions is through regression approach. [8]. Approximate methods however exist including the so-called method of weighted means, assuming all the assumptions for the use of ANOVA t-test are satisfied [6], [10], [1]. We therefore however present an alternative non-parametric method that will take care of different factors and interaction effects.

2. Methodology

Let x_{ilj} be the *ith* observation at the l^{th} level of factor A and j^{th} level of factor B, for $i=1,2,...,nlj$, $l=1,2,...,a$, $j=1,2,...,b$. Let n_{lj} be the number of observations in the $(lj)^{th}$ cell. Then an analysis based on the un-weighted means using the variable [6].

$$\bar{x}_{lj} = \frac{\sum_{i=1}^{n_{lj}} x_{ilj}}{n_{lj}} \quad (1)$$

An analysis of variance is then calculated in the usual way using $\bar{x}_{lj's}$ as if they were single observations for the treatment combination per cell. However, the sum of squares are no longer additive in the sense that the individual sums of squares no longer add up to the total sum of squares. The sum of square error must now be calculated directly and independently from its basic definition, which may be sometimes more time consuming. Hence, instead of using the un-weighted means approach, we will propose a method based on the rank of sample observations.

To develop the proposed method based on the ranks of the sample observations, we would first pool all the $n = \sum_{l=1}^{a}\sum_{j=1}^{b} n_{lj}$ into one commom sample and then rank the observations together in the usual way assingning all the tied observations their mean ranks.

Let r_{ilj} be the rank assigned to x_{ilj} in the combined ranking of these observations, for $i = 1,2,...,lj$, $l = 1,2,...,a$ and $j = 1,2,...,b$, giving a total of n rank with mean ranks $\bar{r} = \frac{n+1}{2}$.

Now, in the absence of ties between the n sample observations, the total sum of squared deviations of the asigned ranks r_{ilj} from their mean rank $\bar{r}$ is given by:

$$SS_{Total} = \sum_{l=1}^{a}\sum_{j=1}^{b}\sum_{i=1}^{n_{lj}}\left(r_{ilj}-\bar{r}\right)^2 = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}$$

$$SS_{Total} = \frac{n(n^2-1)}{12} \quad (2)$$

Which has a chi-square distribution with $n-1$ degrees of freedom.

Now, the total sum of square SS_{Total} can be partioned into two component sums of square, namely, the treatment sum of square SST and the error sum of square SSE as:

$$SS_{Total} = \sum_{l=1}^{a}\sum_{j=1}^{b}\sum_{i=1}^{n_{lj}}\left(r_{ilj}-\bar{r}\right)^2$$

$$= \sum_{l=1}^{a}\sum_{j=1}^{b}\sum_{i=1}^{n_{lj}}\left(\left(\bar{r}_{.lj}-\bar{r}\right)+\left(r_{ilj}-\bar{r}_{.lj}\right)\right)^2$$

where $\bar{r}_{.lj}$ is the mean or average of the ranks assigned to the n_{lj} observations in the l^{th} level of factor A and the j^{th} level of factor B, for $l = 1,2,...,a$, $j = 1,2,...,b$.

It can easily be shown that:

$$SS_{Total} = \sum_{l=1}^{a}\sum_{j=1}^{b}\sum_{i=1}^{n_{lj}}\left(r_{ilj}-\bar{r}\right)^2$$

$$= \sum_{l=1}^{a}\sum_{j=1}^{b} n_{lj}\left(\bar{r}_{.lj}-\bar{r}\right)^2 + \sum_{l=1}^{a}\sum_{j=1}^{b}\sum_{i=1}^{n_{lj}}\left(r_{ilj}-\bar{r}_{.lj}\right)^2$$

$$SS_{Total} = \sum_{l=1}^{a}\sum_{j=1}^{b}\frac{R_{.lj}^2}{n_{lj}} - \frac{n(n+1)^2}{4} + \sum_{l=1}^{a}\sum_{j=1}^{b}\sum_{i=1}^{n_{lj}}\left(r_{ilj}-\bar{r}_{.lj}\right)^2 \quad (3)$$

where $R_{.lj}$ is the sum of the ranks assigned to observations in the l^{th} level of factor A and the j^{th} level of factor B.

Now the sum of squares treatment,

$$SST = \sum_{l=1}^{a}\sum_{j=1}^{b}\frac{R_{.lj}^2}{n_{lj}} - \frac{n(n+1)^2}{4} \quad (4)$$

has chi-square distribution with $a.b-1$ degrees of freedom. (Hogg et al 2005), while the error sum of square

$$SSE = \sum_{l=1}^{a}\sum_{j=1}^{b}\sum_{i=1}^{n_{lj}}\left(r_{ilj}-\bar{r}_{.lj}\right)^2 \quad (5)$$

has chi-square distribution with $(n-1)-(a.b-1) = n-a.b$ degrees of freedom.

The treatment sum of squares SST can be further partitioned in to the sum of squares for factor A, SSA, sum of squares for factor B, SSB, and the factor A by factor B interaction sum of squares $SSAB$. Thus,

$$SST = \sum_{l=1}^{a}\sum_{j=1}^{b} n_{lj}\left(\bar{r}_{.lj}-\bar{r}\right)^2$$

$$= \sum_{l=1}^{a}\sum_{j=1}^{b} n_{lj}\left(\left(\bar{r}_{l.}-\bar{r}\right)+\left(\bar{r}_{.j}-\bar{r}\right)+\left(\bar{r}_{.lj}-\bar{r}_{l.}-\bar{r}_{.j}+\bar{r}\right)\right)^2$$

which can easily be shown to reduce to

$$SST = \sum_{l=1}^{a}\sum_{j=1}^{b}\frac{R_{.lj}^2}{n_{lj}} - \frac{n(n+1)^2}{4}$$

$$SST =$$

$$b\left(\sum_{l=1}^{a}\frac{R_{l.}^2}{n_{l.}} - \frac{n(n+1)^2}{4}\right) + a\left(\sum_{j=1}^{b}\frac{R_{.j}^2}{n_{.j}} - \frac{n(n+1)^2}{4}\right) + \sum_{l=1}^{a}\sum_{j=1}^{b} n_{lj}\left(\bar{r}_{.lj}-\bar{r}_{l.}-\bar{r}_{.j}+\bar{r}\right)^2 \quad (6)$$

That is

$$SST = SSA - SSB - SSAB \quad (7)$$

Where $\bar{r}_{l.}$ and $\bar{r}_{.j}$ are respectively the mean ranks of observations of the l^{th} level of factor A and j^{th} level of factor B, $n_{l.}$ and $n_{.j}$ are the corresponding number of observations and $R_{l.}$ and $R_{.j}$ are the corresponding rank totals, for for $l = 1,2,...,a$, $j = 1,2,...,b$.

Now, the sum of square due to factor A namely:

$$SSA = b\left(\sum_{l=1}^{a}\frac{R_{l.}^2}{n_{l.}} - \frac{n(n+1)^2}{4}\right) \quad (8)$$

has chi-square distribution with $a-1$ degrees of freedom and may be used to test the null hypothesis of no factor A effects.

The sum of square due to factor B namely:

$$SSB = a\left(\sum_{j=1}^{b}\frac{R_{.j}^2}{n_{.j}} - \frac{n(n+1)^2}{4}\right) \quad (9)$$

has the chi-square distribution with $b-1$ degrees of freedom and may be used to test the null hypothesis of no factor B effects.

Similarly, the sum of squares due to factor A by factor B interaction namely

$$SSAB = \sum_{l=1}^{a}\sum_{j=1}^{b} n_{lj}\left(\bar{r}_{.lj} - \bar{r}_{l.} - \bar{r}_{.j} + \bar{r}\right)^2 \quad (10)$$

Or

$$SSAB = SST - SSA - SSB \quad (11)$$

has the chi-square distribution with $(ab-1)-(a-1)-(b-1) = (a-1)(b-1)$ degrees of freedom and may be used to test the null hypothesis of no factor A by factor B interaction effects.

In two factor analysis a null hypothesis which is usually of interest is that there are no treatment effects. If this null hypothesis is rejected, then one may proceed to test the null hypothesis that the effects of each of the factors A and B are zero assuming that the interaction effects have been found not to be statistically significant or that the interactions have been removed by an appropriate data transformation.

The null hypothesis of no treatment effect is tested using the chi-square statistic of Equation (4). The null hypothesis is rejected at the α level of significance if

$$\chi_{SST}^2 \geq \chi_{1-\alpha:ab-1}^2 \quad (12)$$

If this null hypothesis is rejected then we would need to first test the null hypothesis of no significant factor A by factor B interaction effects. This null hypotheis is tested using the chi-square statistic for interaction in Equation (11).

The null hypothesis of no factor A by factor B interaction effect is rejected at the α level of significance if the chi-square value of Equation (10) or (11) is greater than the chi-square critical value with $(a-1)(b-1)$ degrees of freedom. If this null hypothesis is rejected, one may then proceed to test the null hypothesis about factor A and factor B effects using Equations (8) and (9) respectively and rejecting the null hypothesis at a specified α level of significance with $a-1$ and $b-1$ degrees of freedom.

3. Illustration

We shall use the data on final cumulative grade point average (FCGPA) of students who graduated in statistics from a certain University by State of origin for four years. The result is presented in Table 1. [10].

Table 1. *FCGPA of graduating students in statistics for the four years by their state of origin.*

Year of Graduation	A	B	C	D	E	F
2005	1.62, 3.22, 1.74, 3,53, 3.04, 2.07, 3.83, 3.15, 2.38, 2.07	2.33, 4.03, 4.00, 2.23, 1.65, 2.57	3.16, 1.66, 3.88, 3.22, 1.72, 1.68, 2.30, 2.45	4.03, 3.46, 2.36, 3.21, 2.73, 1.69, 2.48	2.99, 3.20, 4.20, 2.82, 2.91, 2.41, 2.35, 2.52	1.96, 2.08, 2.93, 2.43, 2.15, 2.56
2006	4.20, 2.99, 1.63, 2.70, 3.34, 2.44, 2.44, 3.62, 2.93, 1.70, 3.12, 2.33	2.99, 2.33, 4.01, 4.20, 3.32, 2.91, 2.23	2.89, 3.18, 2.62, 3.04, 1.87, 1.52, 2.10	4.66, 3.46, 3.38, 2.24, 1.69, 2.56. 2.62	2.42, 2.36, 4.02, 3.82, 3.01, 2.68, 2.52	2.78, 2.79, 3.02, 1.40, 2.19, 2.13
2007	3.12, 1.92, 1.71, 2.00, 2.33, 2.99, 3.54, 3.62, 3.50	2.63, 2.24, 2.82, 2.06, 1.97, 2.32, 2.72, 2.61	3.21, 2.58, 1.82, 2.67, 2.63, 2.87, 2.23, 2.46	2.04, 2.78, 2.93, 2.63, 3.01, 2.52, 2.46, 2.05, 2.26	2.56, 2.42, 2.40, 2.82 2.63, 1.98, 2.01, 2.00	3.20, 2.82, 1.93, 2.45, 2.03, 3.52, 2.81, 1.48
2008	2.69, 1.38, 3.17, 4.04, 4.11, 3.10, 2.14, 3.10, 2.68, 3.40, 1.95, 2.23, 1.56	1.96, 2.18, 1.32, 2.56, 2.48, 2.76, 2.32	3.06, 2.82, 2.63, 3.52, 1.46, 1.82, 2.22	1.23, 3.02, 2.58, 2.76, 2.43, 1.76, 3.28	1.27, 2.56, 1.82, 1.52, 2.68, 3.12, 2.45	1.80, 2.06, 2.62, 2.24, 2.74, 1.96, 2.48, 3.04

State of Origin

Using the unweighted mean approach, we obtain the entries in Table 2 using Equation (1)

Table 2. *Unweighted mean of the observations.*

Year of Graduation	A	B	C	D	E	F	$t_{l.}$
2005	2.66	1.68	2.56	2.85	2.93	2.35	15.03
2006	2.80	2.80	2.46	2.94	2.98	2.39	16.71
2007	2.75	3.14	2.56	2.52	2.35	2.53	15.13
2008	2.49	2.23	2.51	2.44	2.20	2.37	14.24
$t_{.j}$	10.7	9.47	10.09	10.75	10.46	9.64	61.11

States of Origin

The data in Table 2 are subjected to the standard balanced ANOVA technique without interrraction to obtain the sum of squares and the result of the analysis is presented in Table (3)

Table 3. *The ANOVA Table.*

Source of variation	Sum of squares	Degrees of freedom	Mean Square	F-ratio	P-value
Year	0.3687	3	0.1229	1.43	0.2733
State	0.5353	5	0.1071	1.24	0.3391
Error	1.2934	15	0.0862		
Total	2.1974	23			

From the ANOVA table, the p-value for block (year of graduation) and treatment (state of origin) show that they are

not significant.

The proposed method

Observations are pooled together and assigned ranks. In the presence of tied observations, the mean of their rank are assigned to them. Further, the individual observations are replaced with their ranks and presented in Table 4.

Table 4. *Ranks of individual observations.*

	State Of Origin					
Year of Graduation	**A**	**B**	**C**	**D**	**E**	**F**
2005	11, 161.5, 21, 173, 145, 44.5, 178, 153, 72, 44.5	66.5, 180, 13, 183.5, 55.5, 97	154, 15, 179, 82, 20, 62, 14, 161.5	183.5, 87, 16.5, 70.5, 116, 168.5, 159.5	137.5, 74, 188, 90, 131.5, 69, 157.5, 126	32, 134, 50, 46, 77.5, 94
2006	188, 137.5, 114, 165, 79.5, 79.5, 12, 175.5, 134, 18, 151, 66.5	137.5, 131.5, 181, 164, 55.5, 66.5, 188	130, 8.5, 102, 27, 47, 156, 145	190, 94, 166, 16.5, 102, 168.5, 59	75.5, 111, 182, 140.5, 90, 70.5, 177	120.5, 52, 142.5, 122, 5, 48
2007	151, 28, 36.5, 66.5, 19, 137.5, 174, 175.5, 170	106, 63.5, 126.5, 100, 34, 115, 59, 42.5	159.5, 129, 25, 84.5, 106, 55.5, 98.5, 109	40, 106, 134, 90, 140.5, 47, 84.5, 61, 120.5	94, 35, 36.5, 73, 106, 38, 75.5, 126.5	157.5, 29, 171.5, 7, 39, 124, 126.5, 82
2008	113, 10, 55.5, 155, 4, 186, 185, 49, 148.5, 111, 148.5, 30, 167	32, 94, 3, 118.5, 87, 63.5, 51	147, 171.5 106, 25, 96.5, 53, 126.5	48, 22, 98.5, 77.5, 163, 142.5, 118.5	2, 151, 25, 111, 82, 94, 8.5	23, 32, 102, 145, 117, 87, 42.5, 59,

The ranks in each of the cells are summed to obtain $R_{.lj}$ and they presented in table 5

Table 5. *Sum of the Rank Cell ($R_{.lj}$) Frequency of Observation Per Cell.*

Year of Graduation	**A**	**B**	**C**	**D**	**E**	**F**	$R_{.l.}$
2005	1003.5(10)	595.5(6)	687.5(8)	801.5(7)	974(8)	433.5(6)	4495.5(45)
2006	1314.5(12)	924(7)	615.5(7)	796(7)	846.5(7)	490(6)	4992.5(46)
2007	958(9)	646.5(8)	767(8)	823.5(9)	584.5(8)	736.5(8)	4516(50)
2008	1362.5(13)	449(7)	578.5(7)	670(7)	473.5(7)	607.5(8)	4141(49)
$R_{..j}$	4644.5(44)	2615(28)	2648(30)	3091(30)	2878.5(30)	2267.5(28)	18145(190)

State of Origin.

From Table 5, the chi-square values for the source of variations SS_{Total}, SST, SSA, SSB and SSE were obtained and presented in Table 6

Table 6. *Summary for Chi-Square Values for the Sums of Square their Critical Values, Degrees of Freedom and P-Values.*

Source of Variation	**Chi-Square Statistic**	**Degrees of Freedom**	**Chi-square critical value**	**p-value**
SSTrt	57326.54	23	44.18	0.0000
SSA	15943.87	3	12.838	0.0000
SSB	13716.99	5	16.75	0.0000
SSAB	27665.54	15	32.801	0.0000
SSError	514240.96	166	53.672	
SSTot	571567.5	189	53.672	

4. Conclusion

In this paper, we have proposed a non parametric method for two factor data analysis with unequal cell frequencies and interaction. This was done by using the ranks of the sampled observations to obtain the chi-square statistic for the testing the null hypothesis of no treatment effect and no interaction between factor A and factor B.

Further the application of the proposed method is studied in practice by considering a real life example on students' final cumulative grade point average (FCGPA) and State of Origin of these students. The chi-square test statistic were estimated based on the proposed methods and the result obtained showed better estimates when compared with the method of unweighted mean.

References

[1] Eze, F. C. and Chigbu P. E., (2012), Unbalanced Two-way Random Model with Integer-Valued Degrees of Freedom, Journal of Natural Sciences Research Vol. 2, No 10, pp 100 – 107.

[2] Hogg R. V., Mackean, J. W. and Craig, A. T. (2005), Introduction to Mathematical Statistics, 5th Edition, Prentice Hall New Jersey.

[3] Howell D. C., (2008) The Analysis of Missing Data, in Outhwaite, W and Turner S., Handbook of Social Science Methodology, London, Sage.

[4] Little, R. J. A and Rubin, D. B., (2002), Statistical Analysis with Missing Data, Second Edition, Hoboken, Wiley.

[5] Malwane M. A. and Samaradasa W., (1997) Two-way ANOVA with unequal cell frequencies and Unequal Variances, Statistica Sinica vol. &, 631 – 646.

[6] Montgomery D. C. (2001), Design and Analysis of Experiment, John Wiley and Sons, NewYork, 3rd Edition.

[7] Montgomery, D. C. and Peck, E. A., (1982), Introduction to Linear Regression Analysis, John Wiley New York.

[8] Neter J, Kutner M. A. Nachtshein and Wasserman W. (1996), Applied Linear Statistical Models, Graw Hill, USA.

[9] Oyeka I. C. A (2009), Applied Statistical Methods in Sciences, Norben Avocation Publishers, Enugu, Nigeria.

[10] Oyeka, I. C. A, Uzuke, C. A., obiora-Ilouno, H. O. and Maduakor, C. O., (2012), A Non Parametric Two-way Analysis of Variance with un equal observations per cell, Journal of Nigerian Statistical association, Vol. 24, pp 59 – 66.