Journal of Modern Applied Statistical Methods

Volume 2 | Issue 1 Article 13

5-1-2003

Performing Two-Way Analysis of Variance Under Variance Heterogeneity

Scott J. Richter
University of North Carolina at Greensboro, sjricht2@uncg.edu

Mark E. Payton
Oklahoma State University, mpayton@okstate.edu

Journal of Modern Applied Statistical Methods
May 2003, Vol. 2, No.1, 152-160

1538 – 9472/03/$30.00

Performing Two-Way Analysis of Variance Under Variance Heterogeneity

Scott J. Richter
Department of Mathematical Sciences
University of North Carolina at Greensboro

Mark E. Payton
Department of Statistics
Oklahoma State University

Small sample properties of the method proposed by Brunner et al. (1997) for performing two-way analysis of variance are compared to those of the normal based ANOVA method for factorial arrangements. Different effect sizes, sample sizes, and error structures are utilized in a simulation study to compare type I error rates and power of the two methods. An SAS program is also presented to assist those wishing to implement the Brunner method to real data.

Key words: Factorial arrangement of treatments, heterogeneity of variance

Introduction

Normal theory methods for analysis of variance depend on the assumption of homogeneity of the variance of the error distribution. For a one-way treatment structure, modifications are available when the homogeneity of variance assumption is violated. Milliken and Johnson (1992) suggest a method due to Box (1954) when sample sizes are equal. When samples sizes are unequal, they suggest Welch's (1951) test.

For multifactor layouts, however, there are few options available for testing effects of interaction and main effects. A parametric approach to this problem was presented by Weerahandi (1995), but it requires complex and intensive computing and isn't yet practical for use on real data. Papers by Akritas (1990), Thompson (1991) and Akritas and Arnold (1994) present nonparametric rank test statistics in a multi-way ANOVA setting. One should see Brunner, et al. (1997) for a survey of references relating to this topic.

Scott Richter is an assistant professor in the Mathematical Sciences Department at the University of North Carolina at Greensboro. His email address is sjricht2@uncg.edu. Mark Payton is a professor in the Department of Statistics at Oklahoma State University. His email address is mpayton@okstate.edu.

One method that does not require the equal variance assumption is based on a Wald statistic, which has an asymptotic chi-square distribution. This method tends to reject too frequently under the null hypothesis for small samples. In fact, simulations of Brunner, et al. (1997) show the test to be liberal (by as much as 0.05) for small to moderate sample sizes, and they suggest a small sample improvement over the Wald statistic.

Their approach is to use a generalization of chi-square approximations dating back to Patnaik (1949) and Box (1954). Simulation results indicate that this adjustment greatly improves the performance of the Wald statistic, and is effective for sample sizes as small as n=7 per factor combination. They also point out that for equal sample sizes, their statistic is identical to the classical ANOVA F-statistic, and thus their method can be regarded as a robust extension of the classical ANOVA to heteroscedastic designs. They recommend that their method should always be preferred (even in the homoscedastic case) to the classical ANOVA. However, they do not investigate how the performance of their statistic compares to the ANOVA F-statistic.

In this paper, we present results of a simulation study comparing the performance of the Brunner statistic to the ANOVA F-statistic, make a recommendation for the Brunner statistic for moderate sample sizes ($n \geq 7$), and also present a SAS program (SAS Institute, Cary, N.C.) for implementing the method.

Brunner Method

The method of Brunner et al. (1997) is a small sample adjustment to the well-known Wald statistic, which permits heterogeneous variance but is known to have inflated Type I error rates for small sample sizes. Consider a two-way layout a levels of factor A and b levels of factor B. Assume a set of independent random variables $X_{ij} \sim N(m_i, s_i^2),\ i = 1,..., ab.$

Let $\boldsymbol{\mu} = (m_1, m_2,..., m_{ab})'$ denote the vector containing the $a \bullet b$ population means. Then the hypotheses of no main effects and interaction can be written as

$H_0(A): \mathbf{M}_A\boldsymbol{\mu} = 0$
$H_0(B): \mathbf{M}_B\boldsymbol{\mu} = 0$
$H_0(AB): \mathbf{M}_{AB}\boldsymbol{\mu} = 0$

where

$\mathbf{M}_A = \mathbf{P}_a \otimes \frac{1}{b}\mathbf{J}_b$
$\mathbf{M}_B = \frac{1}{a}\mathbf{J}_a \otimes \mathbf{P}_b$
$\mathbf{M}_{AB} = \mathbf{P}_a \otimes \mathbf{P}_b.$

Here , $\mathbf{P}_c = \mathbf{I}_c - \frac{1}{c}\mathbf{J}_c$, where $\mathbf{I}_c$ is a $c \times c$ identity matrix, $\mathbf{J}_c$ a $c \times c$ matrix of 1's, and the symbol $\otimes$ represents the Kronecker product of the matrices. The vector of observed cell means is denoted by $\overline{\mathbf{X}} = (\overline{X}_1,..., \overline{X}_{ab})'$ and the estimated covariance matrix is given by

$\hat{\mathbf{S}}_N = N \bullet diag\left\{\frac{S_1^2}{n_1},..., \frac{S_{ab}^2}{n_{ab}}\right\}$, where S_i^2 is the i^{th} sample variance and $N = \sum_{i=1}^{ab} n_i$.

For a complete cross-classification, the test statistic is $FB = \frac{N \bullet \overline{\mathbf{X}}' \mathbf{M} \overline{\mathbf{X}}}{\frac{1}{(n-1)} tr\left(\hat{\mathbf{S}}_N\right)}$, which has an approximate F distribution with

$f_{num} = \frac{\frac{1}{(n-1)^2} \bullet \left[tr\left(\hat{\mathbf{S}}_N\right)\right]^2}{tr\left(\mathbf{M}\hat{\mathbf{S}}_N\mathbf{M}\hat{\mathbf{S}}_N\right)}$ numerator and

$f_{den} = \frac{\left[tr\left(\hat{\mathbf{S}}_N\right)\right]^2}{tr\left(\hat{\mathbf{S}}_N^2 \mathbf{?}\right)}$ denominator degrees of freedom, where $\Lambda = diag\left\{\frac{1}{n_1 - 1},..., \frac{1}{n_{ab} - 1}\right\}$ (Brunner, 1997).

Results

A simulation study was performed using SAS version 8.02 for a two-way layout with $a = 4$ and $b = 3$, for various sample sizes. The model used for all simulations was

$$Y_{ijk} = a_i + b_j + ab_{ij} + e_{ijk},$$
$$i = 1,2,3,4,\ j = 1,2,3, \quad .$$
$$k = 1,..., n_{ij},\ e_{ijk} \sim N(0, s_{ij}^2)$$

The classical F test from ANOVA (denoted by F), assuming normality and equal variances, and the adjusted F-test (denoted by FB) of Brunner, et al. (1997) were calculated for 5000 samples and the probabilities of rejection estimated using an $\alpha = 0.05$. Differences in Type I error rates and powers are investigated for different sample sizes, effect sizes, and variance structures.

Case 1: Homogeneous errors, equal sample sizes.

For this case, we let $k = 1,..., n,\ e_{ijk} \sim N(0, s_i^2)$. Table 1 shows nominal Type I error rate for both methods, for various sample sizes. Note that the FB statistic

underestimates the nominal level when n is small, but for sample size as small as n = 7, the nominal rates are comparable to the classical ANOVA test. As sample size increases beyond n = 7, the nominal rate remains stable near the target $\alpha = 0.05$.

Tables 2 and 3 give proportion of rejections when factor A effect is present, and when both main effects are present, respectively, for n = 3 and n = 7. When $n = 3$, the test based on the FB statistic has less power than the F statistic, and underestimates the nominal rate, especially for the test of interaction and when the effect size is small. When $n = 7$, power and nominal rate are very similar, with the exception that the nominal rate for interaction is still a bit too low.

Table 4 shows that when interaction only is present, the FB statistic again has less power for the small sample size case. When the sample size is $n = 7$, power is comparable for both tests, especially when effect sizes are not very small.

Table 1. Proportion of rejections at $\alpha = 0.05$, normally distributed errors, equal variance, based on 5000 samples, no effects present, equal cell sample sizes.

		n					
Test for:	Method	2	3	5	7	10	20
Main Effect A	F	.0492	.0496	.0478	.0482	.0494	.052
	FB	.0130	.0284	.0412	.0448	.0462	.0512
Main Effect B	F	.0466	.0522	.0526	.0530	.052	.0466
	FB	.0142	.0360	.0448	.0502	.0502	.0466
Interaction	F	.0458	.0470	.0474	.0512	.053	.0488
	FB	.0086	.0222	.0326	.0402	.0456	.0462

Table 2. Proportion of rejections at $\alpha = 0.05$, normally distributed errors, equal variance, based on 5000 samples, factor A effect present (a_1=c, a_3=-c), equal cell sample sizes.

		n = 3			*n* = 7		
		c			*c*		
Test for:	Method	.5	1.0	1.5	.5	1.0	1.5
Main Effect A	F	.3446	.9302	1.000	.7530	.9998	1.000
	FB	.2642	.8876	.9992	.7370	.9998	1.000
Main Effect B	F	.0522	.0522	.0522	.0530	.0530	.0530
	FB	.0360	.0360	.0360	.0502	.0502	.0502
Interaction	F	.0470	.0470	.0470	.0512	.0512	.0512
	FB	.0222	.0222	.0222	.0402	.0402	.0402

Table 3. Proportion of rejections at $a = 0.05$, normally distributed errors, equal variance, based on 5000 samples, factor A and B effects present ($a_2=b_1=c$, $a_3=b_2=-c$), equal cell sample sizes.

		n = 3			n = 7		
		c			c		
Test for:	Method	.5	1.0	1.5	.5	1.0	1.5
Main Effect A	F	.3440	.9214	.9998	.7422	1.000	1.000
	FB	.2604	.8780	.9986	.7276	1.000	1.000
Main Effect B	F	.5268	.9902	1.000	.9140	1.000	1.000
	FB	.4576	.9830	1.000	.9100	1.000	1.000
Interaction	F	.0470	.0470	.0470	.0512	.0512	.0512
	FB	.0222	.0222	.0222	.0402	.0402	.0402

Table 4. Proportion of rejections at $a = 0.05$, normally distributed errors, equal variance, based on 5000 samples, interaction effect present ($ab_{11}=ab_{33}=c$, $ab_{13}=ab_{31}=-c$), equal cell sample sizes.

		n = 3			n = 7		
		c			c		
Test for:	Method	.5	1.0	1.5	.5	1.0	1.5
Main Effect A	F	.0496	.0496	.0496	.0482	.0482	.0482
	FB	.0284	.0284	.0284	.0448	.0448	.0448
Main Effect B	F	.0522	.0522	.0522	.0530	.0530	.0530
	FB	.0360	.0360	.0360	.0502	.0502	.0502
Interaction	F	.1584	.5976	.9460	.4276	.9828	1.000
	FB	.0842	.4368	.8734	.3864	.9762	1.000

Case 2: Heterogeneous errors, equal sample sizes.

Here we consider:

$$k = 1,...,n,\ e_{ijk} \sim N(0, s_{ij}^2 = (1 + i * j/2)^2),$$

(errors increasing with the levels of A). Tables 5, 6 and 7 are heterogeneous analogs to Tables 2, 3 and 4, respectively. They compare the tests under variance heterogeneity. Note that the classical F-test shows inflated nominal rates for all effects, with the test for interaction the most inflated. The inflation becomes more severe as the ratio between smallest and largest variances becomes larger. The test using the Box-type adjustment, however, maintains the correct nominal rate in all conditions considered.

Table 5. Proportion of rejections at $\alpha = 0.05$, normally distributed errors with unequal variance (variance increasing with factor A levels, ratio of largest to smallest variance ≈ 10 to 1), based on 5000 samples, factor A effect present ($a_1=c, a_3=-c$), equal cell sample size: $n_i=7$.

		c			
Test for:	Method	0	.5	1.5	2.5
Main Effect A	F	.0592	.1684	.9518	.9998
	FB	.0490	.1384	.9266	.9998
Main Effect B	F	.0564	.0564	.0564	.0564
	FB	.0482	.0482	.0482	.0482
Interaction	F	.0728	.0728	.0728	.0728
	FB	.0486	.0486	.0486	.0496

Table 6. Proportion of rejections at $\alpha = 0.05$, normally distributed errors with unequal variance (variance increasing with factor A levels, ratio of largest to smallest variance ≈ 22 to 1), based on 5000 samples, factor A effect present ($a_1=c, a_3=-c$), equal cell sample size: $n_i=7$.

		c			
Test for:	Method	0	.5	1.5	2.5
Main Effect A	F	.0652	.1008	.5324	.9672
	FB	.0488	.0750	.4408	.9392
Main Effect B	F	.0612	.0612	.0612	.0612
	FB	.0488	.0488	.0488	.0488
Interaction	F	.0824	.0824	.0824	.0824
	FB	.0494	.0494	.0494	.0494

Table 7. Proportion of rejections at $\alpha = 0.05$, normally distributed errors with unequal variance (variance increasing with factor A levels, ratio of largest to smallest variance ≈ 22 to 1), based on 5000 samples, factor A and B effects present ($a_2=b_1=c, a_3=b_2=-c$), equal cell sample size: $n_i=7$.

Test for:	Method	.5	1.5	2.5
Main Effect A	F	.1030	.5234	.9518
	FB	.0784	.4422	.9220
Main Effect B	F	.1228	.7868	.9980
	FB	.1014	.7298	.9962
Interaction	F	.0824	.0824	.0824
	FB	.0494	.0494	.0494

Case 3: Homogeneous errors, unequal sample sizes.

In this case we consider:

$$k = 1,...,n_{ij},\ e_{ijk} \sim N(0,1)\ ,$$

where $n_{1j} = 7,\ n_{2j} = 8,\ n_{3j} = 9,\ n_{4j} = 10$. Here there was little difference in the performance of the two tests (See Tables 8 and 9). The Box-adjusted test showed slightly higher power in some cases.

Case 4: Heterogeneous errors, unequal sample sizes.

Here we consider:

$$k = 1,...,n_{ij},\ e_{ijk} \sim N(0, s_i^{\,2})\ ,$$

with $n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$. When the largest variance was associated with the smallest sample the classical F-test always had inflated nominal Type I error rates (often more than twice the nominal rate) for any effects not present, while the Box-adjusted test maintained expected nominal Type I error rates (See Tables 10, 11 and 12). The classical F-test had greater power for small effect sizes, but the power advantage became negligible as the effect size increased.

Although not shown here, when the largest variance was associated with the largest sample the power of the two tests was essentially equivalent, with the Box-adjusted test often having a slight power advantage. The classical F-test tended to underestimate the Type I error rate for effects not present.

Table 8. Proportion of rejections at $a = 0.05$, normally distributed errors with unequal sample sizes ($n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$) and equal variances, based on 5000 samples, factor A effect present ($a_1=c, a_3=-c$).

		C		
Test for:	Method	0	.5	1.5
Main Effect A	F	.0482	.7962	1.000
	FB	.0500	.8258	1.000
Main Effect B	F	.0518	.0552	.0598
	FB	.0514	.0514	.0514
Interaction	F	.0500	.0502	.0462
	FB	.0414	.0414	.0414

Table 9. Proportion of rejections at $a = 0.05$, normally distributed errors with unequal sample sizes ($n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$) and equal variances, based on 5000 samples, factors A and B effects present ($a_2=b_1=c, a_3=b_2=-c$).

		C	
Test for:	Method	.5	1.5
Main Effect A	F	.8002	1.000
	FB	.8302	1.000
Main Effect B	F	.9596	1.000
	FB	.9564	1.000
Interaction	F	.0498	.0496
	FB	.0420	.0420

Table 10. Proportion of rejections at $a = 0.05$, normally distributed errors with unequal sample sizes ($n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$) and unequal variances ($s^2_{1j} = 10, s^2_{2j} = 5, s^2_{3j} = 2, s^2_{4j} = 1$), based on 5000 samples, factor A effect present (a_1=c, a_3=-c).

		c		
Test for:	Method	0	.5	1.5
Main Effect A	F	.1056	.2902	.9850
	FB	.0476	.1666	.9422
Main Effect B	F	.1000	.1024	.1034
	FB	.0418	.0418	.0418
Interaction	F	.1244	.1246	.1230
	FB	.0494	.0494	.0494

Table 11. Proportion of rejections at $a = 0.05$, normally distributed errors with unequal sample sizes ($n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$) and unequal variances ($s^2_{1j} = 10, s^2_{2j} = 5, s^2_{3j} = 2, s^2_{4j} = 1$), based on 5000 samples, factor A and B effects present (a_2=b_1=c, a_3=b_2=-c).

		C		
Test for:	Method	.5	1.0	1.5
Main Effect A	F	.3070	.8176	.9944
	FB	.1634	.6660	.9788
Main Effect B	F	.4522	.9450	.9992
	FB	.3174	.8852	.9980
Interaction	F	.1242	.1224	.1208
	FB	.0494	.0494	.0494

Table 12. Proportion of rejections at $a = 0.05$, normally distributed errors with unequal sample sizes ($n_{1j} = 7, n_{2j} = 8, n_{3j} = 9, n_{4j} = 10$) and unequal variances ($s^2_{1j} = 10, s^2_{2j} = 5, s^2_{3j} = 2, s^2_{4j} = 1$), based on 5000 samples, interaction effect present (ab_{11}=ab_{33}=c, ab_{13}=ab_{31}=-c).

		C		
Test for:	Method	.5	1.5	2.5
Main Effect A	F	.1060	.1046	.1016
	FB	.0476	.0476	.0476
Main Effect B	F	.1032	.1018	.1026
	FB	.0418	.0418	.0418
Interaction	F	.2128	.8278	.9996
	FB	.0938	.6324	.9898

Conclusion

Based on our results and the results of Brunner, et al. (1997), we agree with those authors that there is no reason to use the classical ANOVA F-test, as long as cell sample size is at least 7. For smaller samples, when the normal theory assumptions hold, we prefer the classical ANOVA F-test, since the FB statistic becomes very conservative in this case. When samples are very small and variances are not equal, the ANOVA test suffers from inflated nominal levels and thus should be used with caution. The FB test, on the other hand, is always conservative in these situations, and thus is a good choice for those concerned mostly with avoiding making Type I errors. The obvious trade-off for small sample sizes, however, is that the FB test is virtually powerless to detect small to moderate effects.

Example 1.

We illustrate the method using an example given in Sokal and Rohlf (1995). The data are from an experiment to examine differences in food consumption when rancid lard was substituted for fresh lard in the diet of rats. The data are classified by fat (fresh, rancid) and gender (male, female). The amount of food eaten (in grams) is given in the following table:

	Fats	
	Fresh	Rancid
Gender		
Male	709	592
	679	538
	699	476
Female	657	508
	594	505
	677	539

A SAS program (available from the first author) was used to compute the p-values for both the ANOVA F-test and the FB test. Since cell sample sizes are equal, values of the F and FB statistics are identical. Notice that although the sample sizes are small (n = 3), there is very little difference between the p-values associated with the two methods, and only a strong effect of gender is evident from the data.

Source of variation	F	p-value	FB	p-value
Fats	2.593	0.146	2.593	0.153
Gender	41.969	<0.001	41.969	<0.001
Fats*Gender	0.630	0.450	0.630	0.454

Example 2.

This example utilizes data presented in Kuehl (2000), page 224. It is a 3x2 factorial experiment involving 3 levels of alcohol and two levels of base. Note that the data are unbalanced in terms of the number of replications per treatment combination.

Because the cell sample sizes are not equal, the calculated test statistics are not the same for the two methods, although the conclusions might be the same for both methods depending upon the level of significance the researcher adopted. The FB statistic gives stronger evidence for effects of interaction and main effects.

	Alcohol		
Base	1	2	3
1	90.7	89.3	89.5
	91.4	88.1	87.6
		90.4	88.3
			90.3
Mean	91.05	89.27	88.93
Std Dev	0.49	1.15	1.21
2	87.3	94.7	93.1
	88.3		90.7
	91.5		91.5
Mean	89.03	94.7	91.77
Std Dev	2.19	---	1.22

Source of variation	F	p-value	FB	p-value
Alcohol	1.931	0.195	4.297	0.053
Base	7.167	0.023	12.858	0.006
Alcohol*Base	7.357	0.011	14.087	0.002

References

Akritas, M.G. (1990), The rank transform method in some two-factor designs. *Journal of the American Statistical Association, 85*, 73-78.

Akritas, M.G., & Arnold, S.F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated-measures designs. *Journal of the American Statistical Association*, *89*, 336-343.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, *25*, 290-302.

Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, *92*, 1494-1502.

Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis*. (2nd ed.) Pacific Grove, CA: Brooks/Cole.

Milliken, G.A., & Johnson, D.E. (1992). *Analysis of messy data, Volume 1: Designed experiments*. New York: Chapman and Hall.

Patnaik, P.B. (1949). The noncentral χ^2 and F-distributions and their applications. *Biometrika*, *36*, 202-232.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry: The principles and practices of statistics in biological research*, New York: W. H. Freeman and Company.

Thompson, G.L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association, 86*, 410-419.

Weerahandi, S. (1995). ANOVA under unequal error variances. *Biometrics, 51*, 589-599.

Welch, B.L. (1951). On the comparison of several mean values. *Biometrika, 38*, 330-336.