Journal of Modern Applied Statistical Methods

Volume 13 | Issue 2 Article 4

11-2014

Conover's F Test as an Alternative to Durbin's Test

Donald J. Best
University of Newcastle, john.best@newcastle.edu.au

John Charles Rayner
University of Newcastle, John.Rayner@newcastle.edu.au

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 76-83.

ISSN 1538 − 9472

Conover's F Test as an Alternative to Durbin's Test

D. J. Best
University of Newcastle
Newcastle, Australia

J. C. W. Rayner
University of Newcastle
Newcastle, Australia

Data consisting of ranks within blocks are considered for balanced incomplete block designs. An F test statistic from ANOVA is better approximated by an F distribution than the Durbin statistic is approximated by a chi-squared distribution. Indicative powers demonstrate that the F test is generally superior to Durbin's test.

Keywords: Analysis of variance, balanced incomplete blocks, F tests, powers, ranks data, taste-tests, test sizes

Introduction

Sometimes the number of treatments to be compared is so large that a complete blocks experiment is impractical. This happens, for example, in some agronomic variety trials. A balanced incomplete block (BIB) design can be used in such a situation. In sensory evaluation trials loss of sensitivity can occur when the subjects are not be able to compare more than a few products with any certainty. Again BIB designs are useful.

Consider a balanced incomplete block design with the data being ranks within blocks. A traditional test for treatment differences for such a scenario is the Durbin (1951) test, based on the statistic D, given by

$$D = c\sum_{i=1}^{t}\left(\bar{R}_i - \frac{k+1}{2}\right)^2$$

Dr. Best is a conjoint Associate Professor in the School of Mathematical and Physical Sciences. Email him at john.best@newcastle.edu.au. Dr. Rayner is a conjoint Professor in the School of Mathematical and Physical Sciences. Email him at john.rayner@newscastle.edu.au.

in which there are t treatments, k of which are ranked in each of b blocks. If r_{ij} is the rank given to treatment i on block j then $\bar{R}_i$ is defined as the mean rank over blocks for the i^{th} treatment. For untied data $c = 12(t-1)r/\{t(k^2-1)\}$ where each treatment is ranked r times, with $r < b$. For tied data, if $V = \sum_{i,j} r_{ij}^2 / (rt) - (k+1)^2 / 4$ then $c = (t-1)r/(tV)$.

It is well known that D has an asymptotic χ^2_{t-1} distribution. However for values of (t, b, k, r) met in practice this approximation to the distribution of D can be poor. See, for example, Fawcett and Salter (1987). This has led to the suggestion to use a permutation test to obtain p-values for D. See, for example, Bi (2009) and Higgins (2004) who does not consider BIBs but who generally advocates permutation tests. However, software to calculate a p-value for D via a permutation test may not always be readily available. A scientist without access to software for a permutation test based on D might find carrying out a permutation test a challenge. Conover (1999, p. 389) suggests carrying out an analysis of variance (ANOVA) on the ranks and using the F test for treatment differences. This F test is based on 'adjusted' sums of squares from the general linear model readily available in statistical packages such as JMP (use 'fit model') and MINITAB (use 'glm'). These packages also readily give appropriate multiple comparisons.

Literature reviews did not reveal any previously published small sample studies examining the validity of this F test approach, although Conover (1999, p. 390) suggested it improves on the Durbin test. As above we observe that we are considering situations where the raw data are ranks or ranks of ordered categorical data. Many studies have compared parametric and nonparametric tests when data are continuous measurements. See, for example, Kelley and Sawilowsky (1997) and the references therein. However, such studies are not the focus of our article.

Sizes and powers

Test sizes based on 100,000 samples for each (t, b, k, r) combination in Table 1 were carried out to check the suitability of the χ^2_{t-1} and $F_{t-1,df}$ distributions for obtaining p-values. Note that, as usual, $df = bk - t - b + 1$. Table 1 sizes are for data with untied ranks with χ^2_{t-1} and $F_{t-1,df}$ critical values and were found using permutation tests. Sizes for D in this study agree with those of Fawcett and Salter (1987). The sizes for the F test statistic F improve on those for D based on the χ^2_{t-1} approximation and indicate the F distribution can be used to obtain p-values for F. If carrying out a permutation test is not convenient the F probabilities are a considerable improvement over the χ^2 probabilities for all bk.

Table 1. Actual test sizes for nominal 5% level test for various BIBs with no ties

(*t*, *b*, *k*, *r*)	*D*	*F*	*bk*
(4, 6, 2, 3)	0.000	0.000	12
(4, 4, 3, 3)	0.000	0.073	12
(5, 10, 2, 4)	0.000	0.116	20
(5, 5, 4, 4)	0.026	0.062	20
(5, 10, 3, 6)	0.035	0.061	30
(6, 15, 2, 5)	0.022	0.050	30
(6, 10, 3, 5)	0.026	0.062	30
(6, 15, 4, 10)	0.030	0.051	60
(6, 20, 3, 10)	0.040	0.056	60
(7, 7, 3, 3)	0.000	0.087	21
(7, 7, 4, 4)	0.025	0.055	28
(7, 21, 2, 6)	0.017	0.069	42

Table 2. Actual test sizes for nominal 5% level test for various BIBs with ties

(*t*, *b*, *k*, *r*)	*D*	*F*	*bk*
(4, 6, 2, 3)	0.000	0.006	12
(4, 4, 3, 3)	0.005	0.051	12
(5, 10, 2, 4)	0.001	0.045	20
(5, 5, 4, 4)	0.018	0.044	20
(5, 10, 3, 6)	0.031	0.052	30
(6, 15, 2, 5)	0.012	0.053	30
(6, 10, 3, 5)	0.024	0.050	30
(6, 15, 4, 10)	0.042	0.050	60
(6, 20, 3, 10)	0.040	0.050	60
(7, 7, 3, 3)	0.001	0.054	21
(7, 7, 4, 4)	0.020	0.050	28
(7, 21, 2, 6)	0.020	0.054	42

To allow for ties, sizes were calculated as in Brockhoff et al. (2004, section 4 and also see the discussion in section 6). For each block and treatment one of the scores 1, 2, …, k was randomly assigned, each with probability $1/k$. These values were ranked by block with ties given mid-rank values. This was repeated 100,000 times for each of the (t, b, k, r) combinations in Table 2. Very infrequently the value V or the error sum of squares was zero. Such data sets were discarded and new ones inserted. Sizes for D are still poor but those for F are better for the ties case than for the no ties case. If, for tied ranks, permutation tests rather than the Monte Carlo tests suggested herein had been used to get sizes for Table 2, results would have been conditional on a ties structure and so not of as general applicability as those given.

A power comparison between the tests is now provided based on D and F. In practice it is expected that most scientists will use the χ^2 and F distributions and so powers based on these are provided. However Tables 1 and 2 show that the test based on D hardly ever has size near 0.05; thus, the D powers based on χ^2 critical points will be disadvantaged in comparison to the F powers based on the F distribution. If test sizes for tied data are examined, it is observed that D sizes for $(t, b, k, r) = (6, 15, 4, 10)$ and $(6, 20, 3, 10)$ are at least not too far from 0.05 and so the D test should not be too disadvantaged. Subsequent powers are calculated following the size method but with all treatments in a given treatment group having probabilities (p_1, p_2, p_3, p_4) of getting a score (1, 2, 3, 4) respectively in any given block instead of (0.25, 0.25, 0.25, 0.25). Thus for $(t, b, k, r) = (6, 15, 4, 10)$ with probabilities (0.25, 0.25, 0.25, 0.25) for treatments 1, 2 and 3, and probabilities (0.08, 0.08, 0.42, 0.42) for treatments 4, 5 and 6 for a nominal 5% level of significance, it was found that the D and F test powers are 0.31 and 0.34 respectively. Recall that under the null hypothesis all treatment probabilities are (0.25, 0.25, 0.25, 0.25). The powers here are close, and the difference could be explained by the discrepancy in the actual sizes. It is expected that - if there was no difference in the sizes - the sizes would be, as here, very close, and there would be no reason, based on power, to use D rather than F. For $(t, b, k, r) = (6, 20, 3, 10)$ and treatment group probabilities as above, then the D and F powers are respectively 0.26 and 0.29. Again, these are very similar and any difference may well be due to the size advantage enjoyed by the F test. It must be stressed that the powers just given are for a BIB design where the actual size was near the nominal size. For the many BIB designs where this is not so, powers of the test based on D would be very poor compared to those of the test based on F.

To further compare the powers of the tests based on D and F and to check whether or not it is the slight size difference that is causing the differences in power, Table 3 gives powers for $(t, b, k, r) = (6, 15, 4, 10)$ for a number of alternative treatment probabilities, using an estimated critical value of 10.64 for the test based on D. Also given are powers using the χ^2_5 critical value of 11.07, which gives a test size of 0.042, whereas 10.64 gives a test size of 0.05.

In all cases the F test power is found to be slightly superior to the Durbin test power; using the estimated critical value of 10.64 it is superior by so little as to be inconsequential. Using the χ^2_5 critical value the difference is small but not inconsequential. Therefore, use of the F test is recommended based on its test sizes being closer to nominal than the Durbin test sizes. Moreover the F test power is generally not inferior, and when the Durbin test has a low size, it is generally

inferior. The F test is easy to use and has ready availability of multiple comparisons in general linear model platforms.

Table 3. Powers for a 5% significance level, (t, b, k, r) = (6, 15, 4, 10), with ties allowed and alternative probabilities as shown

Treatment groups	Alternative probabilities	*D* (11.07)	*D* (10.64)	*F*
(1, 2, 3) (4, 5, 6)	(0.25, 0.25, 0.25, 0.25) (0.15, 0.15, 0.35, 0.35)	0.112	0.129	0.133
(1, 2, 3) (4, 5, 6)	(0.25, 0.25, 0.25, 0.25) (0.1, 0.1, 0.4, 0.4)	0.229	0.258	0.262
(1, 2, 3) (4, 5, 6)	(0.25, 0.25, 0.25, 0.25) (0.05, 0.05, 0.45, 0.45)	0.437	0.473	0.478
(1, 2, 3) (4, 5, 6)	(0.25, 0.25, 0.25, 0.25) (0.02, 0.02, 0.48, 0.48)	0.598	0.634	0.639
(1, 2) (3, 4, 5, 6)	(0.25, 0.25, 0.25, 0.25) (0.15, 0.15, 0.35, 0.35)	0.102	0.119	0.122
(1, 2) (3, 4, 5, 6)	(0.25, 0.25, 0.25, 0.25) (0.1, 0.1, 0.4, 0.4)	0.207	0.233	0.236
(1, 2) (3, 4, 5, 6)	(0.25, 0.25, 0.25, 0.25) (0.05, 0.05, 0.45, 0.45)	0.400	0.433	0.437
(1, 2) (3, 4, 5, 6)	(0.25, 0.25, 0.25, 0.25) (0.02, 0.02, 0.48, 0.48)	0.546	0.578	0.584
(1, 2) (3, 4) (5, 6)	(0.25, 0.25, 0.25, 0.25) (0.1, 0.1, 0.4, 0.4) (0.02, 0.02, 0.48, 0.48)	0.551	0.584	0.588
(1, 2) (3, 4) (5, 6)	(0.25, 0.25, 0.25, 0.25) (0.1, 0.1, 0.4, 0.4) (0.15, 0.15, 0.35, 0.35)	0.175	0.197	0.201

Examples

Ice cream data

Suppose, as in Conover (1999, p. 390) that seven varieties of ice cream are to be compared. Also suppose it is known that tasting more than three ice creams at a time will result in poor responses due to sensory fatigue. The seven ice cream judges are each asked to rank three of the seven varieties. The results are in Table 4.

Table 4 shows that $t = b = 7$, $r = k = 3$ and each variety is compared with every other variety once. This is a balanced incomplete block layout; no ties are observed

and $D = 12$. Using the χ^2_6 approximation the p-value is 0.06 and so with a 5% level of significance it may be concluded that no difference exists in the preference for the seven varieties. However Conover calculates an exact p-value of 0.018. This study calculated $F = 8$ and, using the $F_{6,8}$ distribution, the p-value is 0.005: this is much closer to the exact Conover p-value for D, is easier to calculate and is significant at the 5% level. Knowing that the χ^2 approximation to D is poor, it is necessary to reverse the initial judgement and conclude that varieties are not equally preferred. As here the χ^2 p-values are often too conservative.

Table 4. Rankings of seven ice cream varieties

Judge	Variety						
	1	2	3	4	5	6	7
1	2	3	-	1	-	-	-
2	-	3	1	-	2	-	-
3	-	-	2	1	-	3	-
4	-	-	-	1	2	-	3
5	3	-	-	-	1	2	-
6	-	3	-	-	-	1	2
7	3	-	1	-	-	-	2
Sum	8	9	4	3	5	6	7

Table 5. Rankings for breakfast cereals

Judge	Cereal				
	A	B	C	D	E
1	1.5	1.5	3	-	-
2	1	2.5	-	2.5	-
3	1.5	3	-	-	1.5
4	1	-	2	3	-
5	1.5	-	3	-	1.5
6	1	-	-	3	2
7	-	2	3	1	-
8	-	2.5	2.5	-	1
9	-	3	-	2	1
10	-	-	3	2	1
Sum	7.5	14.5	16.5	13.5	8

Breakfast cereal (tied) data

Kutner et al. (2005, section 28.1) consider a taste-test in which five breakfast cereals ($t = 5$) were scored on a ten point hedonic scale by ten judges ($b = 10$) three at a time ($k = 3$). Each cereal was tasted six times ($r = 6$). The ranked data are shown in Table 5 (note that there are tied ranks).

It was found that $D = 14.92$ with a χ^2_4 p-value of 0.005 and $F = 11.56$ with an $F_{4,16}$ p-value of 0.0001. Using a 5% significance level a decision would be made that there was a difference in the preference ranking of the cereals.

Conclusion

The test based on the ANOVA F statistic F provides an easily applied alternative to Durbin's rank test. The test based on the F statistic has better test sizes than the test based on D, has better power if chi-squared critical values are used for D, and can be calculated using the general linear model software available in many statistical packages, which also readily provide multiple comparisons. Based on the results in this study, it is suggested that, for $bk \geq 50$, the F statistic p-value based on the F distribution can be used rather than p-values from permutation or Monte Carlo tests. For smaller bk the F probabilities are a considerable improvement over the χ^2 probabilities and should be used when carrying out a permutation test is not convenient.

References

Bi, J. (2009). Computer intensive methods for sensory data analysis, exemplified by Durbin's rank test. *Food Quality and Preference*, *20*, 195-202.

Brockhoff, P. B., Best, D. J. and Rayner, J. C. W. (2004). Partitioning Anderson's statistic for tied data. *Journal of Statistical Planning and Inference*, *121*(1), 93-111.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.

Durbin, J. (1951). Incomplete blocks in ranking experiments. *British Journal of Psychology* (Statistical Section), *4*, 85-90.

Fawcett, R. and Salter, K. (1987). Distributional studies and the computer: an analysis of Durbin's rank test. *The American Statistician*, *41*, 81-83.

Higgins, J. J. (2004). *Introduction to Modern Nonparametric Statistics*. Belmont, CA: Duxbury Press.

Kelley, D. L. and Sawilowsky, S. S. (1997). Nonparametric alternatives to the F statistic in analysis of variance. *Journal of Statistical Computing and Simulation, 58*, 343-359.

Kutner, M., Nachtsheim, C., Neter, J. and Li, W. (2005). *Applied linear statistical models* (5th ed.). Boston: McGraw-Hill Irwin.