Comparing Nonparametric Tests of Equality of Means for Randomized Block Designs

Communications in Statistics - Simulation and Computation

D. J. Best

School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia

(John.Best@newcastle.edu.au)

J. C. W. Rayner*

Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia and

School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia

(John.Rayner@newcastle.edu.au)

O. Thas

Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium and

Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia

(Olivier.Thas@UGent.be)

Jan De Neve

Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium

(JanR.DeNeve@UGent.be)

* Corresponding author. Fax +61 2 4921 6898; Telephone +61 2 4921 5737

and David Allingham

Centre for Computer-Assisted Research Mathematics and its Applications, School of Mathematical and Physical Sciences, University of Newcastle, Newcastle, New South Wales, Australia

(David.Allingham@newcastle.edu.au)

Abstract

A number of nonparametric tests are compared empirically for a randomized block layout. We assess tests appropriate for when the data are not consistent with normality or when outliers invalidate traditional ANOVA tests. The objective is to assess, within this setting, tests that use ranks within blocks, the rank transform procedure that ranks the complete sample and continuous analogues of the Cochran-Mantel-Haenszel tests. The usual linear model is assumed, and our primary foci are tests of equality of means and component tests that assess linear and quadratic trends in the means. These tests include the traditional Page and Friedman tests. We conclude that the rank transform tests have competitive power and warrant greater use than is currently apparent.

Key Words: alignment; non-normal errors; nonparametric analysis; randomized blocks layout; simulation study.

1. Introduction

The primary aim of this study is to empirically compare, in the context of the randomized blocks layout, nonparametric tests based on ranking within blocks with nonparametric tests that rank overall. Occasionally it will be possible to rank within blocks but not overall. One interpretation of this is that ranking overall contains more information. Our naïve expectation is that this greater information will result in greater power.

Suppose we have continuous, ranked, or ordered categorical data in a randomized blocks layout. Comparing population treatment means is often accomplished using the analysis of variance (ANOVA), but this analysis is not appropriate if the responses are not normally distributed or if there are data outliers. We discuss three types of analysis that are suitable for continuous data when the ANOVA is not appropriate. The three types of analysis are based on

- the original data
- ranks within blocks
- the rank transform that ranks the complete sample.

These analyses are also suitable for ranked or ordered categorical data. Here though the responses are assumed to be continuous, and so ties are mostly avoided.

The hypotheses of greatest focus here involve tests for equality of means. We are interested in decompositions of these test statistics into tests for linear or trend effects, and tests for quadratic effects. Potentially this could be extended to testing for higher order effects, but in data analysis we combine these into an assessment of residual effects. We could also assess dispersion and higher order effects, and its decomposition into linear, quadratic and other polynomial effects, but we choose to not examine this complexity.

The tests to be compared empirically will be defined in the section 2; they include the traditional Page and Friedman tests, analogues of Cochran-Mantel-Haenszel (CMH) tests and rank transform tests. Section 3 gives a size and power study based on those in Iman et al. (1984) and Kepner and Robinson (1984). In section 4 an example is given to demonstrate the way in which the ranking within blocks approach may be implemented in practice.

2. Definitions

Suppose we have continuous, ranked, or ordered categorical data Y_{ij}, $i = 1, \ldots, t$ and $j = 1, \ldots, b$ in a randomized blocks layout where the subscript i denotes the ith of t treatments and the subscript j denotes the jth of b blocks. Comparing population treatment means is often accomplished using the analysis of variance (ANOVA), but this analysis is not appropriate if the responses y_{ij} are not normally distributed or if there are data outliers. The model for the response Y_{ij} is

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

in which the τ_i are treatment effects with $\sum_i \tau_i = 0$, the β_j are block effects with $\sum_j \beta_j = 0$, μ is an overall mean and the ε_{ij} are independent random variables with mean 0 and variance σ^2. Normality is *not* assumed for the error distribution.

Three classes of test are considered. The first is based on the original data and uses a suggestion of Davis (2002, section 8.7). Davis (2002) suggests constructing sparse t by s contingency tables for the t treatments and s distinct ordered data values. The classical CMH 'mean scores' and 'correlation' statistics can then be calculated. ~~These CMH~~ They give identical

answers to those given by CMH software for the test statistics we now define. The CMH mean scores statistic M_C (M for mean and C for continuous data) is used to test for population mean differences and can be simplified for a randomised block layout. It is defined by

$$M_C = b^2 \sum_{i=1}^{t} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 / V .$$

The linear (trend or CMH correlation) statistic L_C is used to test for an *a priori* ordering of the population means and, after simplification for a randomised block layout, is defined by

$$L_C = b^2 (\sum_{i=1}^{t} \lambda_i \bar{Y}_{i\bullet})^2 / (V \sum_{i=1}^{t} \lambda_i^2).$$

The quantity V in M_C and L_C is the sum of the sample variances for each of the b blocks. This is given by

$$V = \{\sum_{i=1}^{t} \sum_{j=1}^{b} Y_{ij}^2 - t \sum_{j=1}^{b} \bar{Y}_{\bullet j}^2\} / (t-1) .$$

The λ_i needed in L_C are orthogonal linear contrast coefficients given in Appendix 2.

In data analysis we also sometimes test for quadratic or umbrella effects by replacing the λ_i in L_C with orthogonal quadratic contrast coefficients π_i also given in Appendix 2. Such a statistic is applied in section 4 and is defined as

$$Q_C = b^2 (\sum_{i=1}^{t} \pi_i \bar{Y}_{i\bullet})^2 / (V \sum_{i=1}^{t} \pi_i^2) .$$

As noted above Davis (2002, section 8.7) suggests the use of M_C and L_C. However he does does not investigate their performance.

Second, we look at a well-known alternative to the ANOVA, which involves taking ranks within blocks (hence RWB). If the Y_{ij} are ranks or mid-ranks for tied data then M_C reduces to Friedman's statistic, M_{RWB} subsequently, and L_C reduces to Page's statistic, L_{RWB} subsequently. The subscript RWB denotes ranks within blocks.

Although M_{RWB} and L_{RWB} can be calculated from the formulae for M_C and C_C above, we note the slightly simpler formulae

$$M_{RWB} = C^* \sum_{i=1}^{t} \{\bar{r}_{i\bullet} - (t+1)/2\}^2 \text{ and } L_{RWB} = C^* \{\sum_{i=1}^{t} \lambda_i \bar{r}_{i\bullet}\}^2 / \sum_{i=1}^{t} \lambda_i^2$$

in which

$$C^* = b(t-1) / \{\sum_{i,j} r_{ij}^2 / b - t(t+1)^2 / 4\}$$

and r_{ij} is the rank or mid-rank, within blocks, of the y_{ij}. When there are no ties $C^* = 12b/\{t(t+1)\}$. Note that a quadratic test statistic Q_{RWB} may be defined using the quadratic contrast coefficients in Appendix 2.2 instead of the linear contrast coefficients. Thus $Q_{RWB} = C^* \{\sum_{i=1}^{t} \pi_i \bar{r}_{i\bullet}\}^2 / \sum_{i=1}^{t} \pi_i^2$.

As a third alternative to ANOVA we follow Conover (1999, p.419) and Iman et al. (1984), who suggest using a rank transform where the bt observations are not ranked within

blocks but rather overall from 1 to *bt*. The usual ANOVA F tests are then applied to these overall ranks. The rank transform statistic for testing population mean differences is

$$M_{\mathrm{RT}} = b\sum_i (\bar{r}_{i\bullet} - \bar{r}_{\bullet\bullet})^2 / (t-1) / \{\sum_{i,j} (r_{ij} - \bar{r}_{i\bullet} - \bar{r}_{\bullet j} + \bar{r}_{\bullet\bullet})^2 / (b-1)(t-1)\}$$

where r_{ij} is now the overall rank of y_{ij} in the full data set. We also consider the rank transform test based on

$$L_{\mathrm{RT}} = b\{\sum_i \lambda_i \bar{r}_{i\bullet}\}^2 / \{(\sum_i \lambda_i^2)\sum_{i,j} (r_{ij} - \bar{r}_{i\bullet} - \bar{r}_{\bullet j} + \bar{r}_{\bullet\bullet})^2 / (b-1)(t-1)\},$$

analogous to L_{RWB}.

Corresponding to Q_{C} and Q_{RWB} we can define

$$Q_{\mathrm{RT}} = b\{\sum_i \pi_i \bar{r}_{i\bullet}\}^2 / \{(\sum_i \pi_i^2)\sum_{i,j} (r_{ij} - \bar{r}_{i\bullet} - \bar{r}_{\bullet j} + \bar{r}_{\bullet\bullet})^2 / (b-1)(t-1)\}.$$

To illustrate the use of the formulae above consider the data in Table 1 taken from Steele et al. (1997, p.580). There are $t = 6$ treatments and $b = 4$ blocks and the data are oil contents in flax seeds. Steele et al. (1997) rank the data within blocks and find an uncorrected for ties value of Friedman's statistic to be 11.07. Using the χ^2_5 approximation this is (just) not significant at the 5% level. However using ranks as scores, M_{RWB}, Friedman's statistic corrected for ties, takes the value 11.23. This is significant at the 5% level.

Suppose *a priori* we expect an ordering of the population means A ≤ B ≤ …≤ G or, similarly, A ≥ B ≥ …≥ G. Then the Page test is appropriate. The square of the standardized Page statistic uncorrected for ties is 5.44, which, using the χ^2_1 approximation, is significant at the 5%

level. Using the ranks as scores, L_{RWB}, Page's statistic corrected for ties, takes the value 5.52. Again this is significant at the 5% level.

We see no reason to not correct for ties.

If the raw data rather than the ranks are used as scores to find M_C and L_C, we find M_C = 12.32 and L_C = 6.87, which, for these data, have smaller p-values than the tests based on M_{RWB} and L_{RWB}. Note that the test based on L_C is significant at the 1% level whereas the Page test is not.

For the Table 1 data we find M_{RT} = 5.08 and, using the $F_{(t-1),\,(b-1)(t-1)}$ distribution as in Conover (1999, p.370), this has a p-value of 0.006. Thus, for the Table 1 data, M_{RT} is the most sensitive of the three population mean difference test statistics used here. Also L_{RT} = 15.23 with p-value 0.001 using the $F_{1,\,(b-1)(t-1)}$ distribution. Thus L_{RT} is the most sensitive of the three linear trend statistics for the Table 1 data.

The above analyses use χ^2 or F approximations to the distributions of the test statistics. For small b and t, as here, it may be wise to also find p-values using a computer intensive permutation test or the Monte Carlo approach in Best et al. (2012). The Q tests are not significant for the oil content data and are not included in the discussion immediately above.

3. Size and Power Study

The study in this section mainly compares the tests based (i) on L_C, L_{RWB} and L_{RT}, and (ii) on M_C, M_{RWB} and M_{RT}. The Q tests were also included. As we have noted, these tests involve, when the data are continuous, use of the raw data, ranks within blocks of the raw data, and overall ranking of the raw data.

As previously noted, the model adopted for this randomized block layout is $Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$ in which, with the usual constraints, τ_i are treatment effects, β_j are block effects, μ is an overall mean and the ε_{ij} are independent random variables with variance σ^2. The τ_i values in Table 2 were chosen to give a range of powers. In Table 2 we take $\mu = 0$, $\sigma^2 = 1$ and $E[\varepsilon_{ij}] = 0$, and consider three error distributions: normal, Laplace and uniform. Apart from the column pertaing to $M_{RT}(\beta)$ all β_j are taken to be zero. The exception is discussed below.

Table 2 shows sizes and powers for the choices of τ_i shown for $t = 3$, $b = 10$ and $\alpha = 0.05$. The Table 2 critical values, based on the appropriate χ^2 and F distributions of M_C, L_C, M_{RT} and L_{RT}, were found to be 5.99, 3.84, 3.55 and 4.41 respectively. Critical values for the Q tests were also based on asymptotic values. The sizes and powers are based on 100,000 Monte Carlo samples. Critical values for the Friedman (M_{RWB}) and Page (L_{RWB}) tests were based on exact values available in Hollander and Wolfe (1999, Appendices 22 and 23). The usual randomization procedure was used to get sizes very close to 0.050 for the Friedman and Page tests.

The Friedman test has slightly less power than that based on M_C for the normal and uniform errors, but slightly better powers for the Laplace errors. The test based on M_{RT} is best for all error distributions.

The Page test is comparable to that based on L_C for the normal and uniform errors, and superior for the Laplace errors. Over all error distributions the test based on L_{RT} is comparable to the tests based on both L_{RWB} and L_C, although the Page L_{RWB} test is perhaps slightly superior.

Generally we found block effects had little effect on the powers in Table 2. However a reviewer correctly suggested that for the test based on M_{RT} test sizes would increase and powers decrease with increasing block effects. See the columns headed $M_{RT}(\beta)$ in Table 2 for powers based on M_{RT} with block effects $\beta_1 = -100$, $\beta_2 = -75$, …, $\beta_8 = 75$, $\beta_9 = 100$, $\beta_{10} = 0$. These block effects are extremely large, but do illustrate the reviewer's suggestion, although the effects on size and power are not large. As block effects are large we might expect that the power of the

test based on $M_{RT}(\beta)$ would be close to the power of the test based on M_{RWB}, and that is the case. Also note that powers in the M_{RT} column of Table 2, rather than the $M_{RT}(\beta)$ column, are similar to unpresented powers for the Kruskal-Wall test for the Table 2 parameters.

The Q tests only performed well for the last alternative in Table 2. The Q_{RWB} powers might have been better had the test size been closer to 0.05 and had quadratic alternatives (τ_i increasing then decreasing or decreasing then increasing) been chosen. Perhaps the test based on Q_{RT} has the most power.

Simulations in Kepner and Robinson (1984) for L_{RWB} agree with ours. The error distributions, as in Kepner and Robinson (1984), were chosen to represent short, medium and fat tailed distributions.

Further power comparisons of the nonparametric randomized block tests are given in Table 3 where, for consistency with the study of Iman et al. (1984), the model is now taken to be $Y_{ij} = \tau_i + \beta_j + \varepsilon_{ij}$ (as previously but omitting μ). The error distributions are also taken from Iman et al. (1984). However this model does *not* assume that the treatment and block effects sum to zero and that the error distributions have zero mean or that $\sigma^2 = 1$. In part this is necessary as in Table 3 blocks are random effects whereas in Table 2 block effects are fixed. We see this choice of model as, in part, giving a robustness assessment because some of the assumptions underpinning some of the tests are not satisfied. We also note that the sizes and powers of Tables 2 and 3 are representative of other choices of parameters we have explored.

A lognormal alternative was selected for the Table 3 comparisons because this skewed distribution often produces outliers and the Cauchy was selected as a symmetric distribution with very fat tails which also produces outliers. We use $t = 5$, $b = 20$, $\alpha = 0.05$ and 100,000 Monte Carlo samples to determine sizes and powers. For $\alpha = 0.05$ critical values used for M_{RWB}, L_{RWB}, M_C, L_C, M_{RT} and L_{RT} were 2.49, 3.84, 9.48, 3.84, 2.49 and 3.96 respectively. As for Table 2 these critical values are derived from χ^2 and F distributions with appropriate degrees of freedom except

that the Friedman and Page critical values are now also based on asymptotic critical values. The M_{RWB} statistic was transformed as in Conover (1999, p.370) to $(b-1)M_{RWB}/\{b(t-1)-M_{RWB}\}$. This random variable is well approximated by the $F_{(t-1),(b-1)(t-1)}$ distribution, from which critical values were calculated. Tables 2 and 3 indicate that permutation test p-values should be used to check the χ^2 p-values for M_C and L_C.

Asymptotic critical values were used for the Q tests. However for these alternatives the powers were hardly distinguishable from the corresponding test sizes, and so powers for the Q tests have been omitted from Table 3. Again the alternatives are not quadratic in nature.

Table 3 also includes sizes and powers after aligning for blocks and then applying the rank transform to the data. Thus we rank values of $Y_{ij} - \hat{\beta}_j$. These test statistics are denoted by M_{ART} and L_{ART}. These aligned rank tests are not given in Table 2 as block effects were negligible. Although we are concerned here with small sample comparisons we note that Mansouri (1998) has given the asymptotic distribution of M_{ART}.

Iman et al. (1984) noted that if there are no block effects then we essentially have a one-way layout. Unpresented results show that when there are no block effects the test based on M_{RT} had similar powers to the Kruskal-Wallis test. When block effects are large, then ranking within (M_{RWB}) and across blocks (M_{RT}) will give the same test results. For this reason they chose alternatives to give block powers close to 0.5. We repeat their setup.

Our powers in Table 3 can be compared with those for $k = 5$ in Iman et al. (1984, Table 5). In Table 3 for the normal and for the lognormal cases the β_j are distributed as N(0, 0.16) and the ε_{ij} are distributed as N(0, 1). If y_{ij} is a random value for the normal case, a random value in the lognormal case is $\exp(y_{ij})$. As explained by these authors, random uniform variates can be used to generate βs and ε_{ij}s for the Laplace case in Table 3. This also applies for the Cauchy and logistic distributions. For the Laplace case the β_j are $-1.1 \log U$ where U is a random U(0, 1) value and the ε_{ij} are distributed as $-2 \log U$. The sign of ε_{ij} is such that the probability of a

positive is equal to the probability of a negative. For the Cauchy case β_j is distributed as 0.2 * $\tan\{\pi(U^* - 0.5)\}$ and the ε_{ij} are distributed as $\tan\{\pi(U - 0.5)\}$. For the uniform case β_j is distributed as $U(0, 0.43)\}$ and the ε_{ij} are distributed as U. In addition to the Iman et al. (1984) error distributions we also consider a logistic error term with ε_{ij} distributed as $\log\{U/(1 - U)\}$ and β_j distributed as 0.2 * $\log\{ U^*/(1 - U^*)\}$. Note that to obtain the Laplace powers shown in Iman et al. (1984) for $k = 5$ and $b = 20$ the τ_5 value should be 1.8, not 2.8.

From Table 3 M_C and L_C do best for the normal and uniform ε_{ij} distributions. The tests based on M_{RT} and L_{RT} seem best for the lognormal, Laplace and Cauchy ε_{ij} distributions, while the tests based on M_{RWB} and L_{RWB} do well for the lognormal and Cauchy ε_{ij} distributions. The tests based on M_C and L_C have poor sizes and power for the lognormal and Cauchy ε_{ij} values. Thus, as in Table 2, it seems no test is clearly superior when the eight tests are compared. As expected, the two approaches involving ranking the raw data do well for the lognormal and Cauchy cases, for which outliers often occur. The aligned rank transform tests, except for the Cauchy errors, perform very similarly to the rank transform tests. Aligned rank transform tests are usually more effective for analyses involving interactions.

These power comparisons give mixed results. The χ^2 approximation to M_C and L_C was shown to be poor for small samples and the tests based on the rank transformation were marginally better than the rank transform tests that rank within blocks. This seems plausible: ranking overall requires more information and this translates into marginally more power.

4. Extended Analysis for Within Block Rankings

We now illustrate a comprehensive analysis using entrenched tests based on ranking continuous data within blocks. We are not aware that this sort of analysis can be extended to the ranking overall situation. The omnibus Anderson (A) test, which compares barplots or distributions of competing treatments can be given, and as well as the test for mean effects (M_{RWB}). We define the omnibus statistic, A below.

For the oil content data given in Table 1 suppose we randomly break the ties and obtain Table 4 (a). From this table we obtain Table 4 (b). M_{RWB} may be partitioned into the Page statistic, L_{RWB}, the quadratic or umbrella statistic, Q_{RWB}, and a residual.

Following Best (1993) we now give a new definition of M_{RWB} for the case of no tied rankings and define a dispersion sensitive test statistic D_{RWB} for the case of no tied rankings. Notice that D_{RWB} applies only to rankings within blocks and so was not included in the power comparisons. The statistic D_{RWB} assesses differences in variances or dispersion of the treatments. Assuming $t > 2$, calculate the orthonormal polynomials

$$g_1(j) = B^*\{(j-1) - (t-1)/2\} \text{ and } g_2(j) = D^*\{(j-1)^2 - (t-1)(j-1) + (t-1)(t-2)/6\}$$

in which

$$B^* = 2\sqrt{\frac{3}{t^2-1}} \text{ and } D^* = 6\sqrt{\frac{5}{(t^2-1)(t^2-4)}}.$$

Take N_{ij} to be the count in the (i, j)th cell of the $t \times t$ table of counts of treatments by rankings as in Table 4 (b), based on ranks within the b blocks. The mean or location effect for the ith treatment M_i say, is defined to be

$$M_i = \sqrt{\frac{t-1}{bt}} \sum_{j=1}^{t} N_{ij} g_1(j)$$

and the variance or spread effect for the ith treatment, V_i say, is defined to be

$$V_i = \sqrt{\frac{t-1}{bt}} \sum_{j=1}^{t} N_{ij} g_2(j) .$$

Then

$$M_{\text{RWB}} = \sum_{i=1}^{t} M_i^2 \text{ and } D_{\text{RWB}} = \sum_{i=1}^{t} V_i^2 .$$

Suppose now we calculate the usual Pearson X^2 statistic for testing independence for the Table 4 (b) counts. Anderson's statistic is $A = \{(t-1)/t\}X^2$ and has asymptotic distribution $\chi^2_{(t-1)^2}$. The statistic D_{RWB}, like M_{RWB}, has asymptotic distribution $\chi^2_{(t-1)}$.

It is now possible to give an overall analysis like that for continuous data. Table 5 is based on within blocks ranking. For the oil content data the extended analysis of Table 5 is not particularly illuminating, but for other data sets this type of analysis could be quite important. An R package that gives the Table 5 calculations except for Q_{RWB} and D_{RWB} has been given by Allingham and Best (2012).

As above we suppose that there is an ordering of the treatments and so we calculate L_{RWB} and Q_{RWB}. Other orthogonal contrasts may also be appropriate. For example, it may be known a priori that treatment G might give more oil content and so contrast $Z = \bar{r}_{6\bullet} - (\bar{r}_{1\bullet} + \bar{r}_{2\bullet} + \bar{r}_{3\bullet} + \bar{r}_{4\bullet} + \bar{r}_{5\bullet})/5$ would be of interest. Contrast Z is a Meittinen's (1969) contrast.

The Table 5 analysis can be given for tied ranking following Brockhoff et al. (2004). Best et al. (2006) consider the same analysis for tied ranks in balanced incomplete blocks.

5. Conclusion

The power comparisons tended to favour tests based on the rank transform. Perhaps the Friedman and Page tests based on within block ranking had the least impressive powers. However extended analysis is available for within block rankings but not as yet for CMH or rank transform tests. For small samples it is suggested that p-values be based on permutation or other computer intensive methods as well as the asymptotic χ^2 distributions. Use of computer intensive methods is particularly important for the CMH tests.

We would like to thank a reviewer for a number of important insights.

Appendix 1: CMH Analogues

See Rayner and Best (2012) for more detail than is given subsequently.

A1.1 *A Continuous Analogue of the CMH Mean Score Statistic*

The CMH mean score statistic is based on the treatment means $\bar{Y}_{i\bullet}$, obtained by averaging over blocks.

Recall that in section 1 we have assumed the model $Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$. For this model the treatment means over blocks, $\bar{Y}_{i\bullet}$, are, by the Central Limit Theorem, approximately N($\mu + \tau_i$, σ^2/ b) when the number of blocks, b, sufficiently large. We may then apply the result that if

$X_1, \ldots, X_n$ are IN(μ, σ^2) then $\sum_i (X_i - \bar{X})^2 / \sigma^2$ is χ^2_{n-1} distributed. It follows that under the null hypothesis that all $\tau_i = 0$, $b\sum_{i=1}^{t} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 / \sigma^2 = M'$ say, is χ^2_{t-1} distributed. Further routine analysis reveals that when testing $\tau = 0$ against $\tau \neq 0$ the null hypothesis is rejected for large values of M'.

In practice σ^2 is unknown. On the jth block, because the sample variance is an unbiased estimator of the population variance, $\mathrm{E}[\sum_{i=1}^{t} (Y_{ij} - \bar{Y}_{\bullet j})^2 / (t-1)] = \sigma^2$. Writing $V = \{\sum_{i=1}^{t}\sum_{j=1}^{b} Y_{ij}^2 - t\sum_{j=1}^{b} \bar{Y}_{\bullet j}^2\} / (t-1)$ as in section 2, summing over blocks and taking the expectation under the null hypothesis gives $\mathrm{E}[V] = b\sigma^2$. In M' replacing σ^2 by its unbiased estimate V/b gives $M_C = b^2\sum_{i=1}^{t} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 / V$ as a test statistic for testing $\tau = 0$ against $\tau \neq 0$. Following Davis (2002, section 8.7) we take its approximate null distribution as χ^2_{t-1}.

A1.2 *A Continuous Analogue of the CMH Correlation Statistic*

First a contrast that can be the basis of an analogue of the CMH correlation statistic is constructed. If X_i has mean μ_i and standard deviation σ_i then a *contrast* in these random variables is a function $\sum_i a_i X_i$ such that $\mathrm{E}[\sum_i a_i X_i] = 0$ and $\sum_i a_i^2 = 1$. Thus if $\lambda_1, \ldots, \lambda_t$ are such that $\lambda_1 + \ldots + \lambda_t = 0$ then because $\mathrm{E}[\sum_i \lambda_i Y_{i\bullet}] = \mu\sum_i \lambda_i = 0$, $C' = \sum_i \lambda_i Y_{i\bullet} / \sqrt{\sum_i \lambda_i^2}$ is a contrast.

From A1.1 immediately above we know that the $\bar{Y}_{i\bullet}$ are, under the null hypothesis $\tau = 0$, approximately distributed as N(μ, σ^2/b). Hence C' is approximately distributed as N(0, σ^2/b) and $b(C')^2/\sigma^2$ is approximately distributed as χ^2_1. A further approximation is introduced if, again as in A1.1, σ^2 is replaced by its unbiased estimate V/b. Then $b^2(C')^2/V$ □= $b^2(\sum_i \lambda_i Y_{i\bullet})^2 / \{V\sum_i \lambda_i^2$

} = L_C say is approximately distributed as χ_1^2. So L_C is approximately the square of a contrast in the response means with approximate distribution χ_1^2.

The CMH correlation statistic is based on the correlation between the treatment scores and the responses aggregated over blocks. First we construct the sample correlation between the treatment scores $\{\lambda_i\}$ and the response means $\{\bar{Y}_{i\bullet}\}$. The treatments scores are assumed to sum to zero; their sample variance is $\sum_i \lambda_i^2 / t$. The ith treatment has population variance σ^2/b, which is estimated by V/b^2. It is this quantity that is used instead of the sample variance of the $\{\bar{Y}_{i\bullet}\}$ **(not $\{\bar{y}_{i\bullet}\}$)**. The sample correlation between the $\{\lambda_i\}$ and the $\{\bar{Y}_{i\bullet}\}$ is, subject to this adjustment, $\sum_i \lambda_i \bar{Y}_{i\bullet} / \sqrt{\{(\sum_i \lambda_i^2) V / b^2\}}$. The square of the random variable corresponding to this approximate correlation is L_C, a continuous analogue of the CMH correlation statistic.

Appendix 2: Linear and Quadratic Coefficients

A2.1 *Linear Coefficients*

t	$\lambda_1, \lambda_2, \ldots, \lambda_t$	$\sum_{j=1}^{t} \lambda_j^2$
3	– 1, 0, 1	2
4	– 3, – 1, 1, 3	20
5	– 2, – 1, 0, – 1, 2	10
6	– 5, – 3, – 1, 1, 3, 5	70
7	– 3, – 2, – 1, 0, 1, 2, 3	28

A2.2 *Quadratic Coefficients*

t	$\pi_1, \pi_2, \ldots, \pi_t$	$\sum_{j=1}^{t} \pi_j^2$
3	1, – 2, 1	6
4	1, –1, – 1, 1	4
5	2, –1, – 2, – 1, 2	14
6	5, – 1, – 4, – 4, – 1, 5	84
7	5, 0, – 3, – 4, – 3, 0, 5	84

References

Allingham, D. and Best, D.J. (2012). Crblocks: Categorical randomized block data analysis. R package version 0.9-1.

Best (1993). Extended analysis for ranked data. *Australian Journal of Statistics*, 35, 257-262.

Best, D.J., Brockhoff, P.B. and Rayner, J.C.W. (2006). Tests for balanced incomplete block ranked data with ties. *Statistica Neerlandica* 60, 3-11.

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, 93-111.

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

Davis, C.S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer.

Hollander, M. and Wolfe, D.A. (1999). *Nonparametric Statistical Methods* (2nd ed.). Chichester: New York.

Iman, R.L., Hora, S.C. and Conover, W.J. (1984). Comparison of asymptotically distribution-free procedures for the analysis of complete blocks. *Journal of the American Statistical Association*, 79, 674-685.

Kepner, J.L. and Robinson, D.H. (1984). A distribution-free rank test for ordered alternatives in randomized complete block designs. *Journal of the American Statistical Association*, 79, 212-217.

Mansouri, H. (1998). Limiting distribution of the aligned rank transform tests in balanced incomplete blocks, *Journal of Statistical Planning and Inference*, 74, 353-364.

Miettinen, O.S. (1969). Individual matching with multiple controls in the case of all-or-non response. *Biometrics*, 25, 339-355.

Rayner, J.C.W. and Best, D.J. (2012). Continuous analogues of Cochran-Mantel-Haenszel statistics. Centre for Statistical and Survey Methodology, The University of Wollongong, Working Paper 12-4.

Steel, R. G. D., Torrie, J. H. and Dickey, D. A. (1997). *Principles and Procedures of Statistics* (3rd ed.). New York: McGraw Hill.

Table 1

Oil content in Redwing flax seeds

	Treatments					
Blocks	A	B	C	D	E	G
1	4.4	3.3	4.4	6.8	6.3	6.4
2	5.9	1.9	4.0	6.6	4.9	7.3
3	6.0	4.9	4.5	7.0	5.9	7.7
4	4.1	7.1	3.1	6.4	7.1	6.7

Table 2

Sizes and powers of several nonparametric tests with $t = 3$, $b = 10$ and $\alpha = 0.05$ based on 100,000 simulations

(a) Normal errors

τ	L_{RWB}	L_C	L_{RT}	M_{RWB}	M_C	M_{RT}	$M_{RT}(\beta)$	Q_{RWB}	Q_C	Q_{RT}
(3 * 0)	0.049	0.047	0.051	0.049	0.040	0.051	0.052	0.032	0.048	0.052
(–.25, 0, .25)	0.22	0.19	0.18	0.11	0.12	0.14	0.12	0.028	0.040	0.051
(–.5, 0, .5)	0.56	0.56	0.54	0.32	0.40	0.42	0.34	0.022	0.027	0.055
(.25, –.5, .25)	0.04	0.04	0.05	0.25	0.30	0.33	0.27	0.27	0.44	0.46

(b) Laplace errors

τ	L_{RWB}	L_C	L_{RT}	M_{RWB}	M_C	M_{RT}	$M_{RT}(\beta)$	Q_{RWB}	Q_C	Q_{RT}
(3 * 0)	0.049	0.044	0.050	0.049	0.036	0.051	0.055	0.030	0.045	0.052
(–.25, 0, .25)	0.29	0.20	0.24	0.15	0.12	0.18	0.16	0.027	0.039	0.052
(–.5, 0, .5)	0.67	0.57	0.67	0.43	0.43	0.54	0.46	0.018	0.024	0.054
(.25, –.5, .25)	0.04	0.03	0.05	0.34	0.33	0.44	0.36	0.38	0.46	0.56

(c) Uniform errors

τ	L_{RWB}	L_C	L_{RT}	M_{RWB}	M_C	M_{RT}	$M_{RT}(\beta)$	Q_{RWB}	Q_C	Q_{RT}
(3 * 0)	0.049	0.048	0.051	0.049	0.047	0.052	0.055	0.030	0.048	0.051
(–.25, 0, .25)	0.22	0.18	0.17	0.11	0.13	0.13	0.12	0.028	0.042	0.052
(–.5, 0, .5)	0.52	0.52	0.50	0.30	0.38	0.38	0.32	0.023	0.028	0.054
(.25, –.5, .25)	0.04	0.03	0.05	0.23	0.29	0.29	0.25	0.25	0.43	0.40

* Corresponding author. Fax +61 2 4921 6898; Telephone +61 2 4921 5737

Table 3

(a) Sizes of several nonparametric tests based on 100,000 Monte Carlo simulations when $t = 5$, $b = 20$, $\tau_i = 0$ and $\alpha = 0.05$

ε_{ij} distribution	L_{RWB}	L_C	L_{RT}	L_{ART}	M_{RWB}	M_C	M_{RT}	M_{ART}	Q_{RWB}	Q_C	Q_{RT}	Q_{ART}
Normal	0.051	0.050	0.049	0.051	0.052	0.046	0.050	0.049	0.051	0.049	0.050	0.050
Lognormal	0.051	0.044	0.050	0.050	0.051	0.033	0.050	0.050	0.052	0.044	0.050	0.051
Laplace	0.050	0.048	0.049	0.049	0.050	0.044	0.050	0.050	0.052	0.051	0.051	0.051
Uniform	0.051	0.050	0.051	0.049	0.051	0.046	0.051	0.050	0.051	0.049	0.049	0.049
Cauchy	0.051	0.030	0.051	0.051	0.052	0.014	0.051	0.051	0.051	0.026	0.051	0.051
Logistic	0.050	0.049	0.050	0.049	0.050	0.046	0.051	0.051	0.051	0.049	0.050	0.050

(b) Powers of several nonparametric tests based on 100,000 Monte Carlo simulations when $t = 5$, $b = 20$, and $\alpha = 0.05$

ε_{ij} distribution	τ	L_{RWB}	L_C	L_{RT}	L_{ART}	M_{RWB}	M_C	M_{RT}	M_{ART}
Normal	(0, .1, .3, .5, .7)	0.62	0.71	0.69	0.69	0.39	0.48	0.46	0.47
Lognormal	(0, .1, .3, .5, .7)	0.62	0.51	0.69	0.64	0.39	0.25	0.46	0.42
Laplace	(0, .5, 1.0, 1.5, 1.8)	0.70	0.63	0.73	0.71	0.47	0.40	0.51	0.49
Uniform	(0, .05, .12, .17, .23)	0.68	0.80	0.74	0.75	0.45	0.57	0.51	0.52
Cauchy	(0, .4, .9, 1.2, 1.6)	0.75	0.10	0.74	0.46	0.51	0.04	0.52	0.28
Logistic	(0, .1, .3, .5, .7)	0.28	0.28	0.30	0.30	0.15	0.15	0.18	0.17

Table 4

(a) Rankings of Table 1 data with two tied pairs randomly split

	Treatments					
Blocks	A	B	C	D	E	G
1	2	1	3	6	4	5
2	4	1	2	5	3	6
3	4	2	1	5	3	6
4	2	6	1	3	5	4

(b) Counts of rankings for Table 1 data with ties split as in (a) above

	Rankings					
Treatments	1	2	3	4	5	6
A	0	2	0	2	0	0
B	2	1	0	0	0	1
C	2	1	1	0	0	0
D	0	0	1	0	2	1
E	0	0	2	1	1	0
G	0	0	0	1	1	2

Table 5

Further analysis of the Table 1 data using ranks within blocks

Statistic	df	Value	χ^2 p-value	Permutation test p-value
M_{RWB}	5	10.29	0.068	0.046
C_{RWB}	1	5.29	0.021	0.018
Q_{RWB}	1	1.40	0.294	0.324
Residual	3	3.90	0.273	0.283
D_{RWB}	5	7.68	0.175	0.165
Residual	15	9.54	0.848	0.892
A	25	27.51	0.331	0.253