Journal of Statistical Theory and Practice (2022) 16:66
https://doi.org/10.1007/s42519-022-00287-3

ORIGINAL ARTICLE

Nonparametric Analysis of Balanced Incomplete Block Rank Data

G. C. Livingston Jr[2] · J. C. W. Rayner[1,2]

Accepted: 30 July 2022 / Published online: 8 September 2022

Abstract
Traditional nonparametric analysis of balanced incomplete block rank data usually involves the Durbin test. Here we give an alternative adjustment for the Durbin statistic for when there are ties and mid-ranks are used. The adjusted Durbin statistic is shown to be simply related to the ANOVA F statistic on the ranks, so that the corresponding tests are, in a sense, equivalent. This means both are tests of equality of mean treatment ranks. A simulation study compares the size and power performances of the competing tests. We also apply the nonparametric ANOVA methodology to give tests for univariate moment effects and, when treatments are ordered, for bivariate moment effects. The latter includes umbrella tests.

Keywords Cereal example · Durbin test · Generalised correlation · Kruskal–Wallis test · Mid-ranks · Nonparametric ANOVA · Rank transform test · Umbrella test

Mathametics Subject Classification 62G · 62F

1 Introduction

The balanced incomplete block design was proposed by Yates [13] who called these designs symmetrical incomplete randomised block arrangements. Kempthorne [5] notes there is a need for designs in which a small number of treatments need to be compared in small blocks. An example is experiments involving animals when small

✉ J. C. W. Rayner
John.Rayner@newcastle.edu.au

G. C. Livingston Jr
Glen.LivingstonJr@newcastle.edu.au

1 National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, NSW 2522, Australia

2 Centre for Computer-Assisted Research Mathematics and Its Applications, School of Information and Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia

litter size may preclude applying all treatments to a single litter. In a similar vein, [3, 11.1a] note the design is relevant in applications, some of which are given, in which individuals are required to give a rating but doing so becomes increasingly difficult as the number of objects requiring rating increases.

The classic nonparametric analysis of rank data from the balanced incomplete block design (BIBD) uses the Durbin [4] test. It is typically portrayed as a test of equality of treatment distributions. The chi-squared approximation to the distribution of the test statistic is known to be poor; see Best and Rayner [1]. Not all users of the test are aware of the tie adjustments that improve that approximation.

The ANOVA F test on the ranks is known to give a very reasonable approximation to the 0.05 nominal significance level, but this test is regarded as a competing test, not an equivalent one. Here we show that even when there are ties the ANOVA F test on the ranks is an equivalent test to the adjusted Durbin test. Thus both test for equality of treatment mean ranks. A simulation study assesses the performance of the unadjusted Durbin test, the Durbin test adjusted for ties using mid-ranks and the ANOVA F test on the ranks in approximating the nominal 0.05 significance level. One aim of the study is to explore the effects of various degrees of categorisation.

Our empirical analysis also assesses the power of the competing tests here. We need to choose a model to do this, and we choose the parametric model knowing that will disadvantage the nonparametric tests. We find that when the error distribution is, in a sense, not far from normal (we consider the uniform), the parametric F test is the most powerful in the snapshots of the parameter space we consider. However, if a normal error distribution is contaminated with outliers by mixing with distant uniforms, then the adjusted nonparametric tests can outperform the parametric F test.

Nonparametric ANOVA (NP ANOVA) may be used to assess univariate effects beyond mean effects and bivariate moment effects beyond simple correlation effects. This enables a deeper scrutiny of the data than the Durbin test alone gives.

Ties adjustments to the Durbin test statistic are given in Sect. 2 and used in Sect. 3 to derive a relationship between the ANOVA F statistic and the adjusted Durbin statistic. This establishes the equivalence of the two tests. Section 4 presents the power and size studies. NP ANOVA, both unordered and ordered, are described in Sects. 5 and 6, and all tests are applied to a data set in Sect. 7. A brief conclusion follows.

2 Ties adjustments for the Durbin Test Statistic

In the balanced incomplete block design, each of the b blocks contains k experimental units, each of the t treatments appears in r blocks, and every treatment appears with every other treatment precisely λ times. Necessarily

$$k < t, r < b, bk = rt, \text{ and } \lambda(t-1) = r(k-1)$$

Treatments are ranked within each block. For untied data Durbin's statistic, D is given by

$$D = -\frac{3r(t-1)(k+1)}{(k-1)} + \frac{12(t-1)}{bk\left(k^2-1\right)}\sum_{i=1}^{t} R_i^2$$

in which R_i is the sum of the ranks given to treatment i, $i = 1, \ldots, t$.

Subsequently, we will need the identities

$$1 + \ldots + k = k(k+1)/2,$$
$$1^2 + \ldots + k^2 = k(k+1)(2k+1)/6$$
$$k(k+1)(2k+1)/6 - k(k+1)^2/4 = (k-1)k(k+1)/12.$$

If ties occur and ranks are assigned so that the rank sums are not affected, then an adjusted Durbin test statistic is given by

$$D_{\mathrm{A}} = \frac{(t-1)\left\{\sum_i R_i^2 - \frac{rbk(k+1)^2}{4}\right\}}{\sum_{i,j} r_{ij}^2 - \frac{bk(k+1)^2}{4}}$$

in which r_{ij} is the rank of treatment i on block j. If there are no ties then on block j the ranks are 1, 2, …, k. Using the given identities we find $\sum_{i,j} r_{ij}^2 = bk(k+1)(2k+1)/6$ and the denominator in D_{A} is $b(k-1)k(k+1)/12$. With this simplification, D_{A} reduces to D.

Now suppose that ties occur in groups and each group is assigned the mid-rank of the ranks those observations would otherwise have received. This does not affect the rank sums but does reduce the rank variability. Thus the denominator is affected although the numerator is not.

To see this, suppose a group of t observations, that would otherwise have occupied the ranks $h+1, h+2, .., h+t$, are tied. The sum of the squares of these ranks is

$$(h+1)^2 + \ldots + (h+t-1)^2 + (h+t)^2 = t\,h^2 + 2\left\{\sum_{(j=1)}^{t} j\right\}h + \sum_{j=1}^{t} j^2$$
$$= t\,h^2 + t(t+1)h + t(t+1)(2t+1)/6$$

using the well-known formulae for the sum and the sum of the squares of the first t integers. If mid-ranks are assigned the sum of these mid-ranks squared is

$$t\{h + (t+1)/2\}^2 = t\left\{h^2 + (t+1)h + t(t+1)^2/4\right\}$$

and the difference is

$$t(t+1)(2t+1)/6 - t(t+1)^2/4 = \left(t^3 - t\right)/12.$$

The replacement has reduced the sum of squares by $(t^3–t)/12$.

In any ranking, there may be multiple sets of ties. Suppose that the gth contains t_g observations. It follows that when passing from untied to tied data using mid-ranks for p observations, the sum of squares is reduced by the aggregation of the corrections $(t_g^3 - t_g)/12$ for all groups of tied observations. Thus $\{1^2 + \ldots + p^2\}$ becomes $\{1^2 + \ldots + p^2\} - \sum_g \left(t_g^3 - t_g\right)/12$.

Suppose now that on the jth block the gth group is of size $t_{g,j}$. The denominator in D_{A} is the total sum of squares in the ANOVA with ranks as data, and this is

$$\begin{aligned} b\left\{1^2 + \ldots + k^2\right\} - \sum_{g,j}(t_{g,j}^3 - t_{g,j})/12 - bk(k+1)^2/4 \\ = b(k-1)k(k+1)/12 - \sum_g t_g^3 - t_g)/12 \\ = Cb(k-1)k(k+1)/12 \end{aligned}$$

in which

$$C = 1 - \sum_{g,j}(t_{gj}^3 - t_{gj})/\{b(k-1)k(k+1)\}.$$

The adjusted Durbin statistic using mid-ranks, D_{M} says, is thus $D_{\mathrm{M}} = D/C$. The denominator in D_{M} is C times what it would have been had there been no ties while the numerator is unaffected.

For hand calculation, and when there are few ties, this form for the adjusted Durbin statistic is often more convenient than the general form first given.

3 The Relationship Between the Adjusted Durbin Statistic and the ANOVA F Statistic

Here we show that when there are ties, the ANOVA F test statistic, F_{A} say, is related to the adjusted Durbin statistic D_{A} by

$$F_{\mathrm{A}} = \frac{edf\, D_{\mathrm{A}}}{(t-1)\{b(k-1) - D_{\mathrm{A}}\}}$$

in which $edf = bk–b–t + 1$, the error degrees of freedom. This is an algebraic relationship that assumes no model. When there are no ties $D_{\mathrm{A}} = D$ and the relationship reduces to that derived, for example, in [8, Sect. 2.5.3]. The relationship shows that the adjusted Durbin and ANOVA F tests are equivalent tests, in the following sense. Tests based on statistics S and T are equivalent if and only if there exists a 1–1 function f such that $T = f(S)$ so that $S > (<)\, a$ if and only if $T = f(S) > (<) f(a)$. Thus both tests come to the same conclusion provided the critical points are chosen appropriately.

Thus both tests are testing for equality of treatment mean ranks, not merely for equality of treatment distributions. Parallel results hold for the adjusted Kruskal–Wallis and Friedman tests.

Aside. There are different possible ANOVA analyses for balanced incomplete block designs. See, for example, [6, Sect. 8.5]. Note that a balanced incomplete block design can be regarded as a block design with missing data. Such designs are not *orthogonal* and so it is possible to calculate treatment sums of squares both adjusted and not adjusted for blocks. We use the former.

First note that as the block means are the same on every block, there are no block effects, and the block sum of squares is zero.

The observations are r_{ij}, the rank for treatment i on block j. The total sum of squares is $SS_{\text{Tot}} = \sum_{i,j} r_{ij}^2 - r_{..}^2/(bk) = \sum_{i,j} r_{ij}^2 - bk(k+1)^2/4$ since $r_{..} = bk(k+1)/2$.

The adjusted treatment sum of squares is

$$SS_{\text{Treat}} = \frac{k}{\lambda t} \sum_{i=1}^{t} \{R_i - \text{block average across blocks that contain treatment } i\}^2$$

For any given i, there are r blocks containing treatment i and each block total is $k(k+1)/2$, so the block average across blocks that contain treatment i is $r(k+1)/2$. It follows that

$$SS_{\text{Treat}} = \frac{k}{\lambda t} \sum_{i=1}^{t} \left\{ R_i - \frac{r(k+1)}{2} \right\}^2 = \frac{k(k+1)}{12} D.$$

This uses $\sum_i \{R_i - r(k+1)/2\}^2 = rt\left(k^2-1\right)/\{12(t-1)\}$ from the equation for D and both $bk = rt$, and $\lambda(t-1) = r(k-1)$.

The error sum of squares is

$$\begin{aligned} SS_{\text{Error}} &= SS_{\text{Tot}} - SS_{\text{Treat}} - SS_{\text{Blocks}} \\ &= \sum_{i,j} r_{ij}^2 - bk(k+1)^2/4 - \frac{k}{\lambda t} \sum_{i=1}^{t} \left\{ R_i - \frac{r(k+1)}{2} \right\}^2. \end{aligned}$$

From the equation for D_{A}

$$\frac{D_{\text{A}}}{(t-1)} = \frac{\left\{ \sum_i R_i^2 - \frac{rbk(k+1)^2}{4} \right\}}{\sum_{i,j} r_{ij}^2 - \frac{bk(k+1)^2}{4}}.$$

The ANOVA F statistic is thus

$$F_{\text{A}} = \frac{\frac{SS_{\text{Treat}}}{t-1}}{\frac{SS_{\text{Error}}}{edf}} = \left\{ \frac{edf}{(t-1)} \right\} \left\{ \frac{\frac{k}{\lambda t} \sum_{i=1}^{t} \left\{ R_i - \frac{r(k+1)}{2} \right\}^2}{\sum_{i,j} r_{ij}^2 - \frac{bk(k+1)^2}{4} - \frac{k}{\lambda t} \sum_{i=1}^{t} \left\{ R_i - \frac{r(k+1)}{2} \right\}^2} \right\}$$

$$= \left\{\frac{edf}{(t-1)}\right\} \frac{\left\{\frac{\frac{k}{\lambda t}\sum_{i=1}^{t}\left\{R_i - \frac{r(k+1)}{2}\right\}^2}{\sum_{i,j} r_{ij}^2 - \frac{bk(k+1)^2}{4}}\right\}}{\left\{1 - \frac{\frac{k}{\lambda t}\sum_{i=1}^{t}\left\{R_i - \frac{r(k+1)}{2}\right\}^2}{\sum_{i,j} r_{ij}^2 - \frac{bk(k+1)^2}{4}}\right\}}$$

$$= \left\{\frac{edf}{(t-1)}\right\} \frac{\left\{\frac{\frac{k}{\lambda t}D_{\mathrm{A}}}{(t-1)}\right\}}{\left\{1 - \frac{\frac{k}{\lambda t}D_{\mathrm{A}}}{(t-1)}\right\}} = \left\{\frac{edf}{(t-1)}\right\} \frac{D_{\mathrm{A}}}{\left\{\frac{(t-1)\lambda t}{k} - D_{\mathrm{A}}\right\}}$$

$$= \frac{edf\,D_{\mathrm{A}}}{(t-1)\{b(k-1) - D_{\mathrm{A}}\}}$$

as previously indicated, since $bk(k-1) = (t-1)\lambda t$, which follows from the BIBD identities $\lambda(t–1) = r(k–1)$ and $bk = rt$.

As the observations here are ranks and not normally distributed, only approximately does the ANOVA F statistic have the distribution $\mathrm{F}_{t-1,bk-b-t+1}$.

It is well-known that the F test using the $\mathrm{F}_{t-1,bk-b-t+1}$ distribution improves considerably on the Durbin test using the χ^2_{t-1} distribution. See, for example, Best and Rayner [1]. However, the size study in the next section expands on the previous study to better assess the effect of categorisation.

4 Simulation Studies

In this section, we perform two simulation studies. The first to assess the test sizes for the unadjusted Durbin, adjusted Durbin using mid-ranks and ANOVA F tests when testing with a nominal significance level of 0.05. The second looks at the power of each of the tests across several scenarios.

4.1 Size Study

We have established that the Durbin test adjusted for ties by using mid-ranks and ANOVA F test on the ranks are equivalent in the sense that if the critical values are chosen appropriately, both will lead to the same conclusion. However the sampling distributions are different, and an important question is, which more closely approximates an intended 0.05 significance level: the adjusted Durbin test using the asymptotic χ^2 distribution, or the ANOVA F test on the ranks using the F distribution?

Best and Rayner [1] considered the BIBDs $(t, b, k, r) = (4, 6, 2, 3)$, $(4, 4, 3, 3)$, $(5, 10, 2, 4)$, $(5, 5, 4, 4)$, $(5, 10, 3, 6)$, $(6, 15, 2, 5)$, $(6, 10, 3, 5)$, $(6, 15, 4, 10)$, $(6, 20, 3, 10)$, $(7, 7, 3, 3)$, $(7, 7, 4, 4)$ and $(7, 21, 2, 6)$, and to facilitate comparisons we do too. This gives sample sizes bk of between 12 and 60. At the upper end of this range, this should be sufficient for the asymptotic χ^2 distribution for the Durbin statistic to be viable.

A random sample was taken by first generating bk uniform (0, 1) values and categorising these into c intervals of equal lengths. Clearly, $c = 3$ represents a severe categorisation, one that is unlikely to occur in practice. As c increases the categorisation becomes increasing less severe.

The categorised values were randomly allocated consistent with the design and then ranked within blocks, assigning mid-ranks to ties. Then both D and D_{M} were calculated, as was the proportion of values greater than the critical point of the χ^2_{t-1} distribution.

The first entry in each cell uses the unadjusted Durbin test, the second the adjusted Durbin and the third the ANOVA F test.

Similarly, $F_{\mathrm{M}} = edfD_{\mathrm{M}}/\{(t-1)[b(k-1) - D_{\mathrm{M}}]\}$ was calculated as was the proportion of values greater than the critical point of the $\mathrm{F}_{t-1,bk-b-t+1}$ distribution. In all cases, the proportion of rejections in 1,000,000 samples with nominal significance level 0.05 was recorded. The resulting standard error of the estimates is 0.0002.

We emphasise that although the current study uses the same designs as Best and Rayner [1] the treatment of ties is different; we do not use the more complicated approach in Brockhoff et al. [2] that was implemented in Best and Rayner [1]. Moreover here we have 1,000,000 simulations instead of 100,000.

For heavy categorisation and small sample sizes, some values of zero occurred for the Durbin statistics. These reflect samples in which all blocks were uninformative: on each block, all observations were given the same rank. Such samples were discarded and replaced.

The designs in Table 1 are arranged in increasing sample sizes bk. There is a clear improvement in all three tests as bk increases.

In terms of closeness to the nominal significance level uniformly the unadjusted Durbin test is worst, and the ANOVA F test is best.

Apart from the results for $c = 3$, there is a general, although not uniform, consistency across different values of c. Mostly there is an improvement as c increases.

Apart from the (5, 10, 2, 4) design, results are satisfactory for the F test with bk at least 20. For the first three designs in Table 1, the two Durbin tests never reject the null hypothesis! The unadjusted Durbin test is never satisfactory for $c = 3$ and is barely so for $bk = 60$ and the larger c. The adjusted Durbin test is uniformly better than the unadjusted Durbin test but is only really satisfactory for the designs with $bk = 60$.

In the light of these results, we recommend inference be based on the ANOVA F test on the ranks when bk is greater than 20. Otherwise, resampling methods should be used.

Our set-up reflects how data are obtained when judges (blocks) assign scores to treatments, as in a Likert study, and for each judge, scores are ranked. We acknowledge that other set-ups are possible.

4.2 Power Study

Using the same designs above, a power study was performed, but with the level of categorisation limited to $c = 10$ and $c = 100$.

Table 1 Proportion of rejections using the unadjusted Durbin, adjusted Durbin and ANOVA F tests respectively for a nominal significance level of 0.05 based on 1,000,000 simulations

Design (t, b, k, r)	bk	Number of categories (c)				
		$c = 3$	$c = 10$	$c = 20$	$c = 50$	$c = 100$
(4, 6, 2, 3)	12	0.0000	0.0000	0.0000	0.0000	0.0000
		0.0000	0.0000	0.0000	0.0000	0.0000
		0.0197	0.0032	0.0009	0.0001	0.0000
(4, 4, 3, 3)	12	0.0000	0.0000	0.0000	0.0000	0.0000
		0.0022	0.0000	0.0000	0.0000	0.0000
		0.0798	0.0896	0.0879	0.0826	0.0798
(5, 10, 2, 4)	20	0.0000	0.0000	0.0000	0.0000	0.0000
		0.0005	0.0000	0.0000	0.0000	0.0000
		0.0554	0.0995	0.1133	0.1185	0.1198
(5, 5, 4, 4)	20	0.0024	0.0167	0.0217	0.0242	0.0253
		0.0255	0.0269	0.0271	0.0264	0.0265
		0.0612	0.0639	0.0634	0.0629	0.0634
(7, 7, 3, 3)	21	0.0000	0.0000	0.0000	0.0000	0.0000
		0.0000	0.0003	0.0001	0.0000	0.0000
		0.0615	0.0673	0.0734	0.0811	0.0843
(7, 7, 4, 4)	28	0.0014	0.0137	0.0190	0.0225	0.0236
		0.0231	0.0234	0.0244	0.0247	0.0248
		0.0583	0.0589	0.0586	0.0573	0.0568
(5, 10, 3, 6)	30	0.0043	0.0239	0.0303	0.0338	0.0339
		0.0341	0.0358	0.0361	0.0360	0.0350
		0.0575	0.0585	0.0588	0.0597	0.0602
(6, 15, 2, 5)	30	0.0002	0.0071	0.0128	0.0178	0.0199
		0.0144	0.0186	0.0201	0.0210	0.0216
		0.0638	0.0589	0.0551	0.0526	0.0521
(6, 10, 3, 5)	30	0.0018	0.0163	0.0217	0.0250	0.0256
		0.0276	0.0288	0.0282	0.0274	0.0269
		0.0582	0.0611	0.0620	0.0631	0.0626
(7, 21, 2, 3)	42	0.0004	0.0088	0.0133	0.0157	0.0166
		0.0217	0.0242	0.0218	0.0194	0.0185
		0.0587	0.0608	0.0647	0.0681	0.0684
(6, 15, 4, 10)	60	0.0078	0.0298	0.0359	0.0392	0.0402
		0.0418	0.0425	0.0424	0.0418	0.0416
		0.0529	0.0536	0.0535	0.0528	0.0528
(6, 20, 3, 10)	60	0.0052	0.0266	0.0331	0.0371	0.0383
		0.0405	0.0412	0.0412	0.0410	0.0407
		0.0532	0.0536	0.0537	0.0543	0.0552

We required a procedure for generating treatment and block effects that would give one-dimensional snapshot of the parameter space.

First, a set of treatment effect vectors were defined for the different designs. These were $(-4, 0, 2, 2)$, $(-3, -2, 0, 1, 4)$, $(-3, -3, 0, 1, 2, 3)$, $(-3, -2, -2, 0, 1, 3, 3)$ for designs where $t = 4, 5, 6$, and 7 respectively. A multiplier was then applied to this which ranged from 0 to 2 in order to induce a continuum of treatment effects. A

multiplier of 0 corresponds to a calculation of test size. Larger values of the multiplier correspond to larger differences to detect; these should correspond to greater power.

Second, a block effect was applied. These were simply defined as a linear interpolation from -1 to 1 based on the number of blocks. For example, when $b = 5$, the block effects were $(-1, -0.5, 0, 0.5, 1)$.

Lastly, a random error was added. For this, two distributions were used: a uniform distribution ranging from -4 to 4; and a mixture distribution. This distribution was a mixture of the standard normal distribution (90%) and uniform distributions from -7 to -6 (5%), and from 6 to 7 (5%). The idea was to mimic normal errors with outliers, a situation in which ranks are likely to be appropriate.

For a particular combination of inputs, 1,000,000 simulations were performed with the number of rejections at the 0.05 level calculated for the unadjusted Durbin, adjusted Durbin using mid-ranks, ANOVA F test on the ranks, and parametric ANOVAs on both the continuous response variable as well as the categorised response variable.

For the uniform error distribution, the results are as one may expect. The ANOVA on the continuous and categorised response performs best followed by the ANOVA on the ranks, then the adjusted and unadjusted Durbin test. As for categorisation, when $c = 10$ the adjusted Durbin test performs much better than the unadjusted; this effect is less marked when $c = 100$.

Where the error distribution is the mixture distribution, the results are a little more varied. There are occasions when the ANOVA on the ranks outperforms both of the parametric ANOVAs and other occasions that show the ANOVA on the ranks outperforming the parametric ANOVA up to a point, after which the parametric tests have superior power. This often occurred around 80% power, sometimes slightly earlier.

The plots in Fig. 1 are representative of all the plots. We see that.

- The difference in the two parametric tests is almost negligible, as is apparent in the differences in the right-hand panels.
- The difference in the powers, as indicated by the right panels, can be in excess of 0.2, and hence the differences between the tests can be substantial.
- For the uniform error distribution, the right-hand panel has curves that are almost always positive, indicating the parametric test on the continuous response is almost uniformly superior.
- The plot with mixture errors and $c = 100$ and hence virtually continuous data, sees the ANOVA on the mid-ranks is superior until overtaken by the parametric test on the continuous response at around 80% power. Even then the advantage of the parametric test is not great.
- The two plots with $c = 10$ and mixture errors show the ANOVA on the mid-ranks is generally superior, with the adjusted Durbin test next best.

These observations lead to the following conclusions. If the error distribution is sufficiently close to normal – as, for example, the uniform is – then the parametric F test is generally superior. There is no need to use the rank tests. However, it is not difficult to construct an error distribution such as the mixture here for which the ANOVA F test on the adjusted ranks may be considerably superior to the parametric tests.

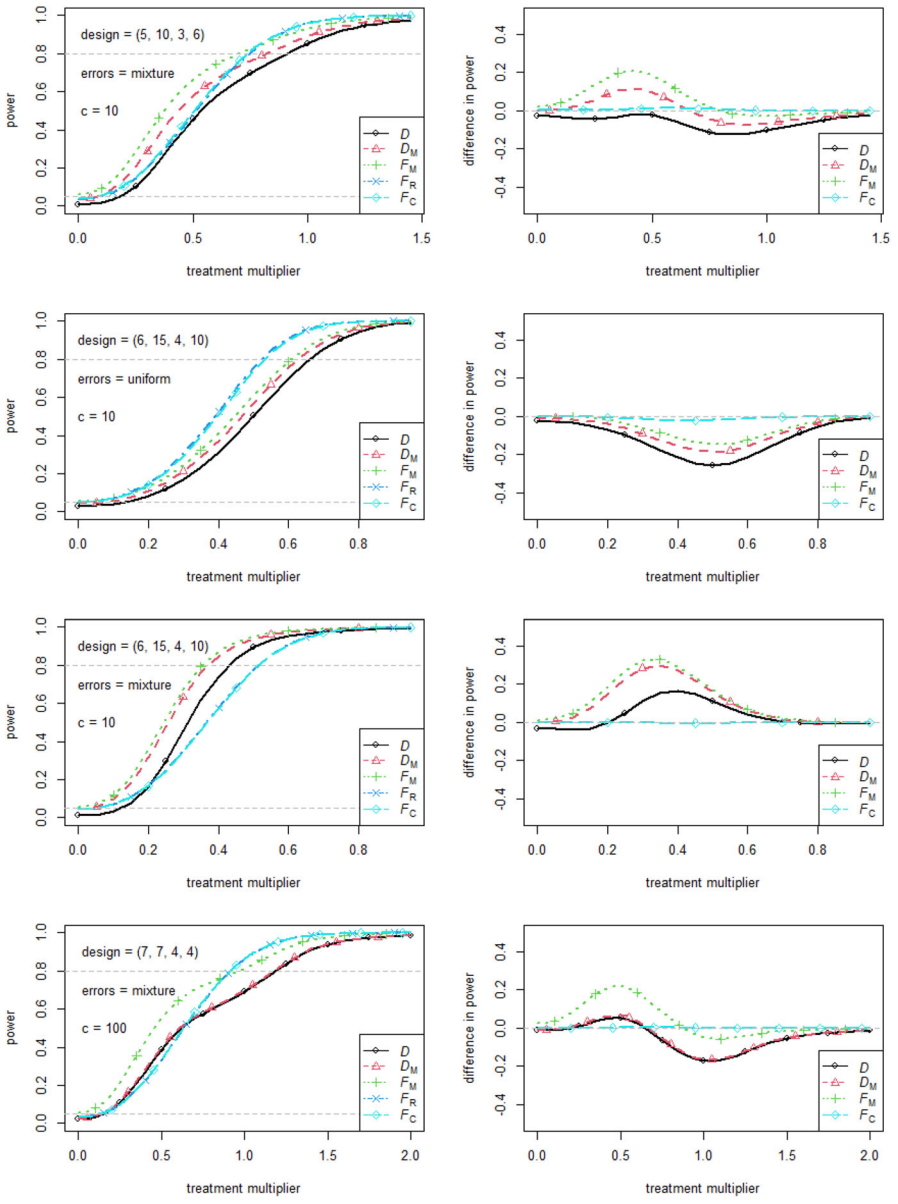

Fig. 1 Power curves for the unadjusted Durbin (D), the adjusted Durbin using mid-ranks (D_{M}), the ANOVA on the mid-ranks (F_{M}), the parametric ANOVA on the continuous response (F_{R}) and the parametric ANOVA on the categorised response (F_{C}). The error distributions, either uniform or mixture, are indicated in the left panel. The plots on the right show the difference in power between each test and the parametric ANOVA on the continuous response

Size considerations indicate the unadjusted Durbin test cannot be recommended. Moreover, the adjusted Durbin test generally had test size slightly less than 0.05 while the ANOVA F test has size slightly greater than 0.05. Between the two tests, the small power advantage for the ANOVA F test may be a reflection of this size advantage.

It would appear that the ANOVA F test is somewhat robust, and if the error distribution is sufficiently close to normal, its use can be recommended. However, if that is not the case, the ANOVA F test on the mid-ranks is to be preferred. The power differences between these two tests can exceed 0.2, so careful consideration of the error distribution is important.

5 Unordered Nonparametric ANOVA

In this and the next section, we use nonparametric analysis of variance (NP ANOVA) to generate a suite of tests for the analysis of BIBD data. This methodology was developed by Rayner and Best [11] and Rayner, Best and Thas [12] in the context of multifactor ANOVA and extended to designs consistent with the general linear model in Rayner [9].

The unordered NP ANOVA requires the use of orthonormal functions. For a random variable X $\{a_u(X)\}$ is a set of orthonormal functions means that

$$\mathrm{E}[a_u(X)a_v(X)] = 1 \text{ if } u = v \text{ and } = 0 \text{ otherwise.}$$

Our preference is to use orthonormal polynomials as they permit moment interpretations. In most circumstances those up to degree three are sufficient. Write μ for the mean of X and μ_r, $r = 2, 3$, for the central moments of X. To give the orthonormal polynomials up to degree r requires moments up to order $2r$. To avoid ambiguity we take $a_0(x) = 1$ for all x. Then

$$a_1(x) = (x-\mu)/\sqrt{\mu_2},$$

$$a_2(x) = \left\{(x-\mu)^2 - \mu_3(x-\mu)/\mu_2 - \mu_2\right\}/\sqrt{d}$$

$$\text{in which } d = \mu_4 - \mu_3^2/\mu_2 - \mu_2^2, \text{ and}$$

$$a_3(x) = \left\{(x-\mu)^3 - a(x-\mu)^2 - b(x-\mu) - c\right\}/\sqrt{e},$$

$$\text{in which } a = (\mu_5 - \mu_3\mu_4/\mu_2 - \mu_2\mu_3)/d,$$

$$b = \left(\mu_4^2/\mu_2 - \mu_2\mu_4 - \mu_3\mu_5/\mu_2 + \mu_3^2\right)/d,$$

$$c = \left(2\mu_3\mu_4 - \mu_3^3/\mu_2 - \mu_2\mu_5\right)/d \text{ and}$$

$$e = \mu_6 - 2a\mu_5 + \left(a^2 - 2b\right)\mu_4 + 2(ab-c)\mu_3 + \left(b^2 + 2ac\right)\mu_2 + c^2$$

The NP ANOVA ignoring order applies the general linear model platform to the data transformed by the orthonormal polynomials of degrees one, two and so on. In

general, there seems to be little to be gained from considering degrees greater than three. The transformations could be of the raw data, the ranked data, or the data scored in some other way. Of course ranking in the context of the BIBD means ranking within blocks.

Because $a_1(x) = (x-\mu)/\sqrt{\mu_2}$ and the ANOVA is location-scale invariant, the analysis of the raw data is identical to the unordered degree one NP ANOVA analysis. Similarly for the analysis of the ranked data, the analysis of the ranked data is identical to the unordered degree one NP ANOVA analysis. This analysis is also the rank transform analysis.

As we are proposing performing several analyses on the same data set, this approach would usually be regarded as exploratory data analysis. That being the case, even if the assumptions underpinning the ANOVA are marginal, we usually nevertheless only consider the ANOVA F test p-values. We then take comfort in the knowledge that ANOVA is broadly robust. Where we have calculated both the F test and resampling p-values we find these are generally in good agreement.

It is also worth mentioning that even if all null hypotheses are true, in performing several tests at the 5% level of significance, roughly 5% of them will reject the null hypothesis. This should be born in mind in interpreting the testing.

6 Ordered Nonparametric ANOVA

If treatments are ordered as well as responses, it is of interest to assess if the two are correlated in the usual linear–linear sense, and also if there are 'umbrella' effects in which as we progress through treatments in a given order, responses either increase then decrease or decrease then increase.

To assess these effects we need the concept of a *generalised correlation*. For bivariate discrete random variables (X, Y) first suppose that $\{a_u(X)\}$ are orthonormal polynomials on $\{p_{i.}\}$ and $\{b_v(Y)\}$ are orthonormal polynomials on $\{p_{.j}\}$. As before, we take the degree zero polynomial to be identically one. For $u \geq 1$ and $v \geq 1$ the (u, v)th generalised correlation is defined to be $\mathrm{E}[a_u(X)\, b_v(Y)]$. The (1, 1) generalised correlation is the 'usual' correlation, while the (1, 2) generalised correlation assesses 'umbrella' effects in which as the second variable increases the first variable increases then decreases or conversely decreases then increase. One of the important properties of generalised correlations is that all the generalised correlations being zero is a necessary and sufficient condition for independence. For more detail on generalised correlations see Rayner and Beh [10].

Given a set of responses $\{y_j\}$ in the ANOVA setting here calculate $\{a_u(y_j)\, b_v(y_j)\}$. To test whether or not the (u, v)th generalised correlation is zero use either the t-test if the data are consistent with normality or a nonparametric test such as the signed-rank test if they are not. Again in the setting here we are particularly interested in assessing linear–linear and umbrella effects.

In ordered NP ANOVA we may also wish to assess whether or not a particular generalised correlation varies across the levels of an unordered factor: here this factor is blocks. To do this assessment two sets of orthonormal polynomials are constructed, on both the responses and the treatments. New responses for a modified design are

then constructed. One such response is the product of the uth orthonormal polynomial on the response and the vth orthonormal polynomial on the treatment. The modified design here is the intended BIBD modified by omitting the treatments since these have been accounted for in the new response. This design is a one-factor ANOVA in which the factor is blocks.

7 Breakfast Cereal Example

The data in Table 2 are from [7, Sect. 28.1]. Each of five breakfast cereals is ranked by ten judges, who each taste three cereals. Each cereal is assessed six times. Thus $t = 5$, $b = 10$, $k = 3$ and $r = 6$. The ANOVA F test on the ranks has F statistic 11.5556 with p-value 0.0001.

For these data $\sum_{i,j} r_{ij}^2 = 137.5$ and $\sum_i R_i^2 = 785$ and using the direct formula $D_{\mathrm{A}} = 14.8571$. Alternatively, the uncorrected Durbin statistic is $D = 13$ and $\sum_{g,j} \left(t_{gj}^3 - t_{gj}\right) = 30$, since in five blocks there are two ties, and in the remaining five blocks there are none. Thus $C = 7/8$ and $D_{\mathrm{M}} = D/C = 13{*}8/7 = 14.8571$, as before. The χ_4^2 p-value for the adjusted Durbin statistics is 0.0050 whereas that for D is 0.0113. The values for F and D_{M} are consistent with the relationship given in Sect. 3.

The first, second and third degree NP ANOVA unordered p-values are respectively 0.0001, 0.8962 and 0.9322 respectively. The corresponding Shapiro–Wilk p-values for the residuals for first, second and third degree analyses are 0.3255, 0.0321, 0.0075 respectively. The first-degree p-value is, as expected, the same as the ANOVA F p-value. There is a strong first-degree effect but no evidence of a higher degree effect, even if the Shapiro–Wilk p-values for the degree two and three analyses make that inference dubious.

Table 2 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

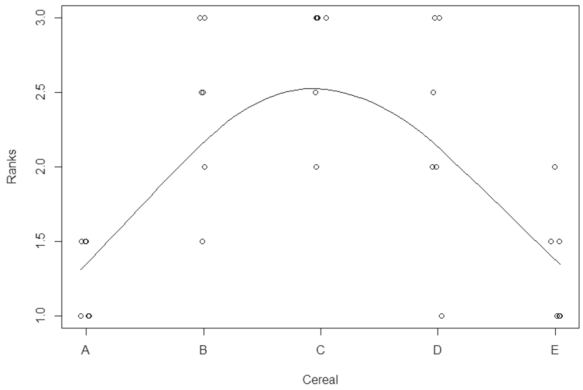

Fig. 2 Mean ranks versus breakfast cereals

We now assume that the cereals are ordered A < B < C < D < E, perhaps by sugar content. The only significant generalised correlation is that of order (1, 2) with both t-test and signed-rank test p-values less than 0.0001. The Shapiro–Wilk test on the residuals has p-value 0.0027, so the t-test result would normally be put aside.

The mean rankings for cereals A to E are 1.1, 2.5, 2.9, 2.3 and 1.2 respectively. A plot of response against cereal in Fig. 2 shows a clear parabolic shape. If the ordering is based on sugar content, then tasters preferred mid-range sugar content.

For this design ordered NP ANOVA involves using the observations of the various generalised correlations as responses and blocks as factor. The only significant result was for the (2, 3)th generalised correlation, indicating that these observations varied across judges. This apparent significance could be a consequence of the fact that when performing multiple tests of significance at the 0.05 level, approximately 5% of the results will be significant under the null hypothesis of no effect. As all the other tests are not significant it appears that the judges are acting homogeneously in their assessments of the cereals.

8 Conclusion

We have shown that when ties occur in the BIBD, whether or not mid-ranks are used, the adjusted Durbin test is equivalent to the ANOVA F test in the sense that each test statistic is a 1 – 1 function of the other. Hence both will reach the same conclusion provided the critical points are chosen appropriately. They therefore are testing the same null and alternative hypotheses, which shows that the adjusted Durbin test assesses equality of mean treatment ranks rather than equality of treatment distributions.

When compared with the Durbin test, adjusted or not, the ANOVA F test is better at approximating the nominal significance level. The adjusted Durbin test is preferred to the unadjusted Durbin test.

In terms of power, the parametric F test on the raw data can have substantially greater power than the nonparametric rank tests we considered if the data are 'sufficiently close' to normal. However, the converse may be true when that is not the case. We

therefore, recommend careful consideration of the error distributions when analysing BIB data.

Nonparametric ANOVA is useful for assessing higher degree univariate and bivariate moment effects.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflict of interest On behalf of both authors, the corresponding author states that there is no conflict of interest.

References

1. Best DJ, Rayner JCW (2014) Conover's F test as an alternative to Durbin's test. J Mod Appl Stat Methods 13:76–83
2. Brockhoff PB, Best DJ, Rayner JCW (2004) Partitioning Anderson's statistic for tied data. J Statist Plann Infer 121(1):93–111
3. Cochran WG, Cox GM (1957) Experimental designs. Wiley, New York
4. Durbin J (1951) Incomplete blocks in ranking experiments. Br J Psychol 4:85–90
5. Kempthorne O (1958) The design and analysis of experiments. Wiley, New York
6. Kuehl RO (2000). *Design of experiments: statistical principles of research design and analysis*. Pacific Grove, CA: Duxbury Press.
7. Kutner M, Nachtsheim C, Neter J, Li W (2005) Applied linear statistical models, 5th edn. McGraw-Hill Irwin, Boston
8. Rayner JCW (2016) Introductory nonparametrics. Bookboon, Copenhagen
9. Rayner JCW (2017) Extended ANOVA. J Stat Theory Pract 11(1):208–219
10. Rayner JCW, Beh EJ (2009) Towards a better understanding of correlation. Stat Neerl 63(3):324–333
11. Rayner JCW, Best DJ (2013) Extended ANOVA and rank transform procedures. Aust NZ J Stat 55(3):305–319
12. Rayner JCW, Best DJ, Thas O (2015) Extended analysis of at least partially ordered multi-factor ANOVA. Aust NZ J Stat 57(2):211–224
13. Yates F (1936) Incomplete randomised blocks. Ann Eugen 7:121–140