2011

Nonparametric tests for latin squares

John Best
University of Newcastle

John Rayner
University of Wollongong

Nonparametric Tests for Latin Squares

D.J. Best
School of Mathematical and Physical Sciences,
University of Newcastle, NSW 2308, Australia
John.Best@newcastle.edu.au and

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

Abstract

A number of nonparametric tests for the Latin square are examined. The rank transform method has good test sizes and powers for the 5 × 5 Latin square for various parameter values and error distributions. Alignment procedures are also examined and their use illustrated using data for replicated Latin squares.

Keywords: Aligned data, Kruskal-Wallis statistic, Rank transform statistic, replicated Latin square.

1. Introduction

Nonparametric tests for some of the simpler experimental designs are well known. Three of the best known are the Kruskal-Wallis test for the one-way layout, the Friedman test for randomised blocks and the Durbin test for the balanced incomplete block design. The test statistics for these tests are commonly given in textbooks such as Higgins (2004) or in software packages. For more complicated experimental layouts there are no such well known tests, but general nonparametric approaches such as (i) the rank transform and (ii) ranking after alignment methods have been suggested. Here we compare (i) and (ii) when applied to Latin square experimental designs.

A Latin square experimental design is often used where there are two blocking factors. As is common these factors will be called rows and columns. If there are t products to compare, each product occurs once in each row and column. The $t \times t$ Latin square is an incomplete three way factorial design with one observation per cell. Only t^2 cells are needed to evaluate the effect of products, rows and columns, whereas the three way factorial with one observation per cell needs t^3 cells. We now give an example.

Burns Example.

Box, Hunter and Hunter (2005, p.170) consider the following.

> Six burn treatments, A, B, C, D, E and F, were tested on six subjects (volunteers). Each subject has six sites on which a burn could be applied for testing (each arm with two below the elbow and one above). A standard burn was administered at each site and the six treatments were arranged so that each treatment occurred once with every subject once in every position. After treatment each burn was covered by clean gauze; treatment C was a control with clean gauze but without treatment. The data are the number of hours for a clearly defined degree of partial healing to occur.

Table 1 gives the data while Figure 1 shows the value 100 in the fifth row and third column is a possible outlier. Hence it may be more appropriate to use a nonparametric analysis than a parametric analysis. Table 2 gives some results in which *F* is the usual ANOVA (analysis of variance) statistic and the other statistics are defined in section 2 below.

Table 1. Burn data.

	Volunteers					
Position on arm	1	2	3	4	5	6
I	A 32	B 40	C 72	D 43	E 35	F 50
II	B 29	A 37	F 59	E 53	D 32	C 53
III	C 40	D 56	A 53	B 48	F 37	E 43
IV	D 29	F 59	E 67	A 56	C 38	B 42
V	E 28	C 50	B 100	F 46	A 29	D 56
VI	F 37	E 42	D 67	C 50	B 33	A 48

Four possible nonparametric statistics for the Latin square are the

- Kruskal-Wallis (*KW*) statistic ignoring row and column effects,
- Kruskal-Wallis (*AKW*) statistic that adjusts for row and column effects,
- rank transform (*RTF*) statistic and
- aligned data rank transform *F* statistic (*ARTF*) which adjusts for row and column effects.

For each of the tests based on these statistics p-values can be found using the asymptotic χ^2 or F distributions or Monte Carlo simulation.

Another nonparametric approach we do not examine here is the use of permutation testing. Permutation tests are not often available in the commonly available software packages unless additional programming is done. The RTF and ARTF tests defined in section 2 and which we recommend subsequently can be

carried out with no additional programming for the RTF and just a little extra programming for the ARTF. In Table 2 and later the Monte Carlo p-values we give are based on random permutations of the ranks data.

The ARTF test is included in the comparisons of section 3 because its use is suggested by, among others, Higgins (2004, p.310) when there are interaction terms in the statistical model. Section 5 looks at replicated Latin squares in which an interaction term is usually part of the model. The KW and AKW tests are included in section 3, even though extra programming is needed to obtain reasonable p-values, as they closely related to a general nonparametric approach introduced in Rayner and Best (2011).

Figure 1. Burn Data

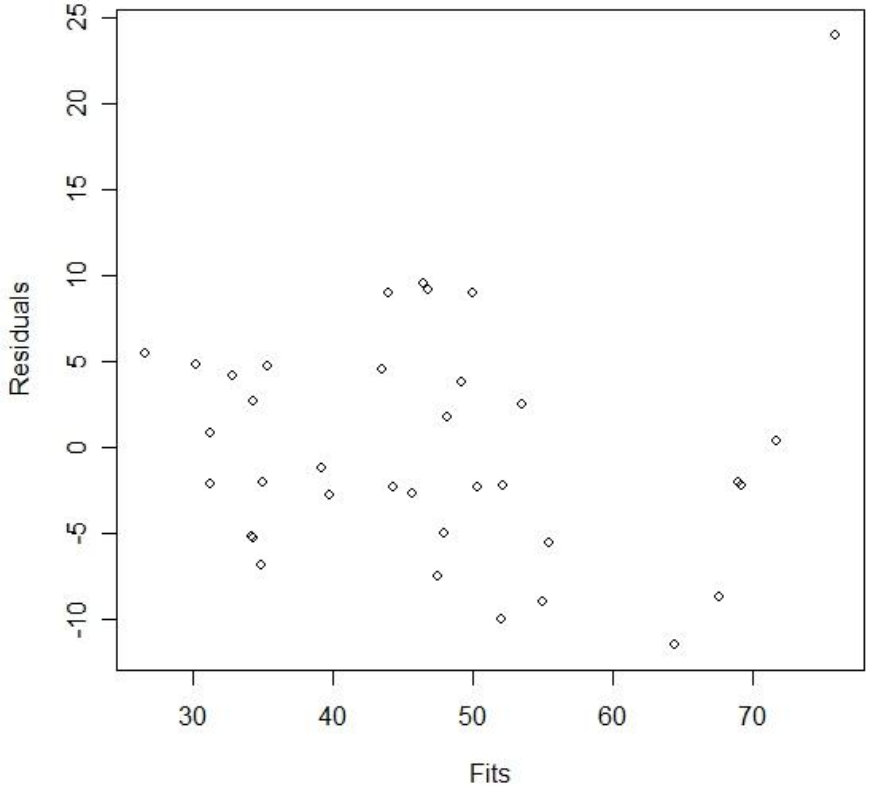


Table 2. Analysis of Burn data.

Statistic	Value	Asymptotic p-value	Monte Carlo p-value
KW	1.989	0.851	0.869
AKW	5.375	0.372	0.385
RTF	1.573	0.213	0.214
ARFT	0.766	0.585	0.579
F	0.585	0.711	0.728

In Table 2 we note the good agreement between the asymptotic and Monte Carlo p-values. This will not always be the case for the KW and AKW tests. The Monte Carlo p-value for the ANOVA F test in Table 2 is based on a permutation test. Notice the wide spread of p-values. In section 4 we give an example where some p-values are below 0.05 and others above. In section 3 we give a small size and power study.

2. Definitions

Following, for example, Kuehl (2000, p.281), a model for Latin square data is

$$Y_{ij} = \mu + \alpha_k + \beta_i + \gamma_j + E_{ij}$$

for product k in row i and column j, where if there are t treatments or products to compare, $i, j, k = 1, ..., t$. Note that if any two of treatments, rows and columns are specified then the design specifies the other product or block. Hence it is equally valid to use any of the notations Y_{ij}, Y_{ijk} and $Y_{ij(k)}$ (and similarly for E_{ij}). The E_{ij} are mutually independent N(0, σ^2) random variables, μ is an overall mean effect, and α_k, β_i and γ_j are parameters that sum to zero, representing fixed treatment (product), row (block) and column (block) effects respectively. A conventional parametric test for differences in product effects is based on an ANOVA F test that is invalid if, for example, the E_{ij} normality assumption does not hold or there are outliers as was the case with the burns example in the Introduction.

To calculate the *KW* statistic the data y_{ij} are ranked from smallest to largest giving ranks r_{ij} say, where tied ranks are given an average rank. Put

- $r_{ij(k)} = r_{ij}$ when product k occupies the (i, j)th cell and zero otherwise,
- $\bar{R}_k = \sum_{i=1}^{t}\sum_{j=1}^{t} r_{ij(k)} / t$ and
- $V = \sum_{i=1}^{t}\sum_{j=1}^{t} r_{ij(k)}^2 / (t^2 - 1) - t(t^2 + 1)^2/\{4(t-1)\}$.

The statistic *KW* is given by

$$KW = \sum_{k=1}^{t} t\{\bar{R}_k - (t^2+1)/2\}/V .$$

To calculate the *RTF* statistic the r_{ij} are subjected to the usual parametric ANOVA and *RTF* is taken to be the ANOVA *F* statistic for between product differences.

To calculate the *ARTF* statistic first align the y_{ij} to obtain

$$y_{ij}^{*} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} - \bar{y}_{..} \text{ in which}$$

$$\bar{y}_{i.} = \sum_{j=1}^{t} y_{ij}/t,\ \bar{y}_{.j} = \sum_{i=1}^{t} y_{ij}/t \text{ and } \bar{y}_{..} = \sum_{i=1}^{t}\sum_{j=1}^{t} y_{ij}/t .$$

The y_{ij}^* are ranked, giving r_{ij}^*. The usual parametric ANOVA is carried out on the r_{ij}^* and *ARTF* is taken to be this ANOVA *F* statistic for between product differences. This alignment can useful if the model is, in fact, linear. If a different model is suspected, a different alignment may be more beneficial.

To calculate the adjusted Kruskal-Wallis test statistic *AKW* use the r_{ij}^* rather than the r_{ij} in the formula for *KW*.

Section 3 considers the 5 × 5 Latin square.

3. Small Size and Power Study

The size study displayed in Table 3 compares actual and nominal test sizes. Generally asymptotic critical values are used; $\chi^2_{4,0.95}$ = 9.4877 for the KW tests and $F_{4,12,0.95}$ = 3.2592 for the F tests. These critical values are used, as, we suggest, this is what practitioners generally use. The RTF test gives sizes fractionally bigger than the nominal value while the ARFT test also has sizes a little greater than nominal. The χ^2 approximation to the KW critical value gives sizes a little on the small side when there are no row or column effects. However when there are row and column effects the actual size of the KW test is much smaller than the nominal size, particularly for the symmetric short tailed U(0, 1) alternative. Thus, as expected, the KW test suffers because the row and column effects are not accounted for. The ANOVA F test has small actual sizes for the exponential and t_2 alternatives and slightly large actual size for the U(0, 1) alternative. As expected it is less distribution free than the RTF or ARFT tests. The χ^2_4 approximation to the AKW critical values is poor and so to apply this test Monte Carlo methods are needed. Some may consider this a disadvantage for the use of the AKW test.

In Table 3 parts (a), (b) and (c) the AKW (1) critical values use the χ^2_4 critical value while those for AKW (2) use 12.5 as the critical value. This value was determined by Monte Carlo methods because the χ^2_4 critical value was inadequate. For other sample sizes and dimensions of the Latin square Monte Carlo would again be needed to determine an adequate critical value. Clearly the AKW (2) sizes here are better.

Table 4 giving powers for the alternatives shown is presented below. There are three treatment or product alternatives given in parts (i), (ii) and (iii) of Table 4 and there are three different combinations of row and column effects given in parts (a), (b) and (c) of Table 4. The AKW values in Table 3 (d) and Table 4 use the 12.5 critical value; for the other statistics the χ^2_4 or $F_{4,12}$ critical values are used.

Table 4 shows, as expected, that the KW test has poor power compared to the other tests when there are row and column effects: see (i) (b), (ii) (b) and (iii) (b). The F, ARFT and AKW tests have less power than the RTF and KW tests when there is an outlier: see (i) (c), (ii) (c) and (iii) (c). This is particularly the case for the U(0, 1) errors which is where the test sizes for the F, ARFT and AKW tests are less than they should be; see Table 3 (d). Perhaps an alignment procedure based on

medians rather than means would help here. Table 4 (ii) (c) shows a distinct divide. Overall the RTF test does well.

Table 3. Test sizes for a sample size of 25 and a nominal significance level of 5%, based on 100,000 Monte Carlo simulations for various parameter configurations.

(a) $\alpha_k = (0, 0, 0, 0, 0)$, $\beta_i = \gamma_j = (0.2, -0.2, 0, 0.2, -0.2)$

Error Distribution	RTF	KW	AKW (1)	AKW (2)	F	ARFT
Normal	0.052	0.027	0.160	0.050	0.050	0.056
Exponential	0.052	0.025	0.139	0.038	0.042	0.051
U(0, 1)	0.053	0.002	0.162	0.051	0.055	0.051
t_2	0.053	0.031	0.135	0.035	0.033	0.050

(b) $\alpha_k = (0, 0, 0, 0, 0)$, $\beta_i = (0.2, 0, 0, -0.2, 0)$, $\gamma_j = (0, -0.2, 0, 0, 0.2)$

Error Distribution	RTF	KW	AKW (1)	AKW (2)	F	ARFT
Normal	0.053	0.031	0.161	0.050	0.050	0.055
Exponential	0.052	0.026	0.140	0.039	0.041	0.051
U(0, 1)	0.053	0.009	0.162	0.053	0.055	0.058
t_2	0.053	0.034	0.135	0.036	0.033	0.052

(c) $\alpha_k = \beta_i = \gamma_j = (0, 0, 0, 0, 0)$

Error Distribution	RTF	KW	AKW (1)	AKW (2)	F	ARFT
Normal	0.053	0.038	0.161	0.050	0.050	0.057
Exponential	0.053	0.040	0.139	0.039	0.041	0.049
U(0, 1)	0.052	0.036	0.163	0.052	0.053	0.056
t_2	0.052	0.036	0.133	0.036	0.033	0.050

(d) $\alpha_k = \beta_i = \gamma_j = (0, 0, 0, 0, 0)$ with outlier of 5.0 at cell (5, 5)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.053	0.036	0.033	0.031	0.051
Exponential	0.053	0.036	0.023	0.030	0.045
U(0, 1)	0.052	0.034	0.004	0.000	0.037
t_2	0.053	0.035	0.033	0.033	0.050

It is interesting to note in parts (a) and (b) of Table 4 that even when there are normal errors the nonparametric tests RTF, ARTF and AKW do as well as the ANOVA F test. In Table 4 (iii) the powers for the U(0, 1) alternative when $\alpha_k = (0.5, -0.5, 0, 0.5, -0.5)$ were all 1.0. If this effect is halved to $\alpha_k = (0.25, -0.25, 0, 0.25, -0.25)$, a more interesting comparison can be made.

The non-normal error distributions used in Table 4 comprised a skewed, a symmetric short-tailed and a symmetric long-tailed distribution. Other choices from these three categories could be made or entirely different error distributions such as bimodal distributions could have been considered. We consider those used as good representatives of their categories. The choice of alternatives and row/column effects is also somewhat arbitrary; for our choices all powers are not all zero or all one. Similar choices for error distributions and alternatives have been made before. See for example Kepner and Robinson (1984).

Table 4. Test powers for a sample size of 25 and a nominal significance level of 5%, based on 100,000 Monte Carlo simulations for various parameter configurations.

(i) (a) α_k = (– 0.5, – 0.25, 0, 0.25, 0.5), $\beta_i = \gamma_j$ = (0, 0, 0, 0, 0)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.19	0.16	0.17	0.19	0.19
Exponential	0.34	0.31	0.22	0.21	0.26
U(0, 1)	0.99	0.99	0.99	0.99	0.99
t_2	0.13	0.09	0.07	0.07	0.10

(i) (b) α_k = (– 0.5, – 0.25, 0, 0.25, 0.5), $\beta_i = \gamma_j$ = (0.5, –0.5, 0, 0.5, – 0.5)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.18	0.05	0.19	0.19	0.19
Exponential	0.26	0.07	0.24	0.24	0.25
U(0, 1)	0.94	0.05	0.99	0.99	0.99
t_2	0.10	0.06	0.06	0.06	0.09

(i) (c) α_k = (– 0.5, – 0.25, 0, 0.25, 0.5), $\beta_i = \gamma_j$ = (0, 0, 0, 0, 0) outlier of 5.0 at cell (5, 5)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.22	0.19	0.16	0.13	0.20
Exponential	0.37	0.35	0.17	0.13	0.24
U(0, 1)	0.99	0.99	0.44	0.73	0.82
t_2	0.14	0.11	0.08	0.07	0.10

(ii) (a) α_k = (0.25, 0, – 0.5, 0, 0.25), $\beta_i = \gamma_j$ = (0, 0, 0, 0, 0)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.12	0.10	0.12	0.12	0.12
Exponential	0.23	0.21	0.15	0.14	0.17
U(0, 1)	0.85	0.85	0.82	0.90	0.85
t_2	0.09	0.07	0.06	0.05	0.08

(ii) (b) α_k = (0.25, 0, – 0.5, 0, 0.25), $\beta_i = \gamma_j$ = (0.5, –0.5, 0, 0.5, – 0.5)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.14	0.03	0.12	0.14	0.13
Exponential	0.18	0.02	0.14	0.14	0.17
U(0, 1)	0.82	0.00	0.82	0.90	0.85
t_2	0.09	0.03	0.06	0.05	0.08

(ii) (c) α_k = (0.25, 0, – 0.5, 0, 0.25), $\beta_i = \gamma_j$ = (0, 0, 0, 0, 0) outlier of 5.0 at cell (5, 5)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.12	0.09	0.07	0.06	0.09
Exponential	0.21	0.19	0.06	0.05	0.10
U(0, 1)	0.80	0.81	0.02	0.00	0.10
t_2	0.09	0.06	0.05	0.05	0.07

(iii) (a) α_k = (0.5, – 0.5, 0, 0.5, – 0.5), $\beta_i = \gamma_j$ = (0, 0, 0, 0, 0)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.29	0.26	0.28	0.28	0.30
Exponential	0.48	0.47	0.36	0.33	0.40
U(0, 1)*	0.68	0.67	0.71	0.71	0.73
t_2	0.17	0.14	0.10	0.09	0.13

(iii) (b) α_k = (0.5, – 0.5, 0, 0.5, – 0.5), $\beta_i = \gamma_j$ = (0.5, – 0.5, 0, 0.5, – 0.5)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.27	0.09	0.28	0.28	0.28
Exponential	0.37	0.11	0.36	0.32	0.38
U(0, 1)*	0.39	0.00	0.72	0.72	0.73
t_2	0.15	0.07	0.10	0.09	0.13

(iii) (c) α_k = (0.5, – 0.5, 0, 0.5, – 0.5), $\beta_i = \gamma_j$ = (0, 0, 0, 0, 0) outlier of 5.0 at cell (5, 5)

Error Distribution	RTF	KW	AKW	F	ARFT
Normal	0.35	0.32	0.28	0.25	0.35
Exponential	0.54	0.53	0.32	0.25	0.42
U(0, 1)*	0.73	0.73	0.14	0.01	0.56
t_2	0.21	0.17	0.12	0.12	0.17

* α_k = (0.25, – 0.25, 0, 0.25, – 0.25) for the U(0, 1) alternative; see text.

4. Traffic Example

This example is chosen to highlight how a different choice of statistical method can result in quite different p-values. Kuehl (2000, p.301) considers the following scenario.

Traffic Example.

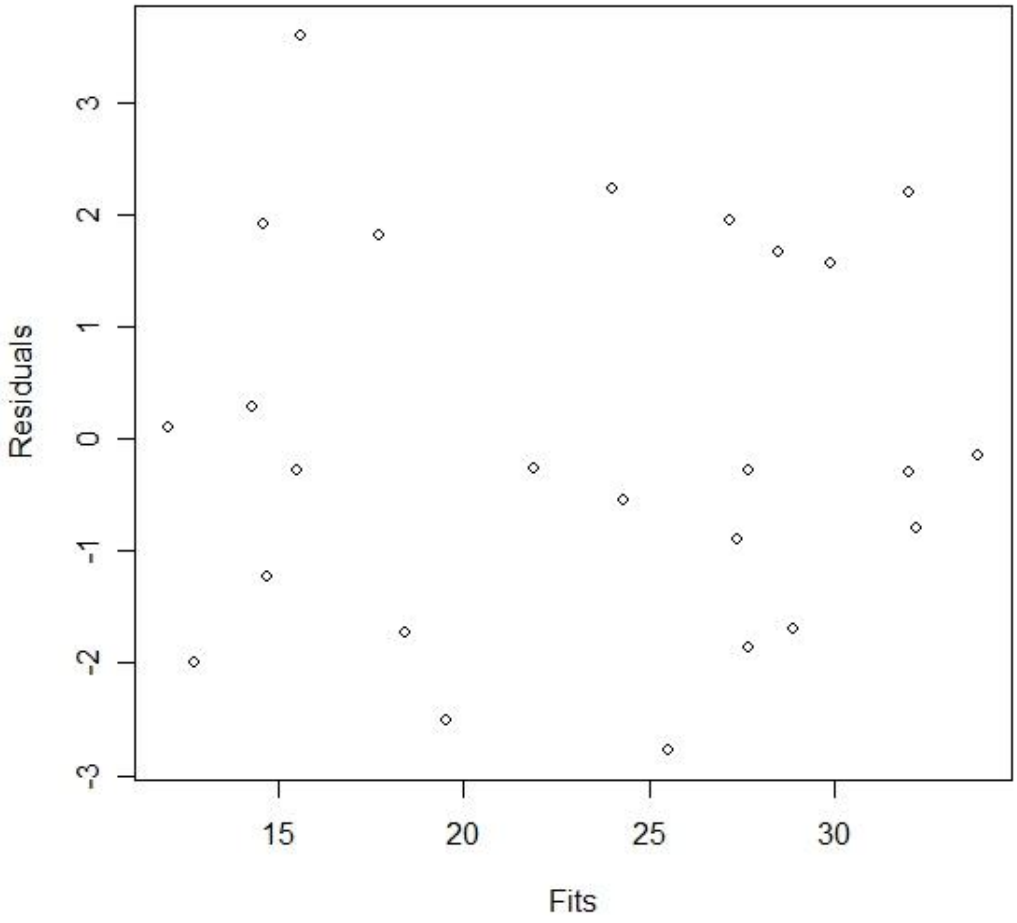
A traffic engineer conducted a study to compare the total unused red light time for five different traffic light signal sequences. The experiment was conducted with a Latin square design in which blocking factors were (1) five intersections and (2) five time of day periods. In Table 5 the five signal sequence treatments are shown in parentheses as A, B, C, D, E and the numerical values are the unused red light times in minutes.

Table 5. Unused red light time in minutes.

	Time Period				
Intersection	1	2	3	4	5
1	15.2 (A)	33.8 (B)	13.5 (C)	27.4 (D)	29.1 (E)
2	16.5 (B)	26.5 (C)	19.2 (D)	25.8 (E)	22.7 (A)
3	12.1 (C)	31.4 (D)	17.0 (E)	31.5 (A)	30.2 (B)
4	10.7 (D)	34.2 (E)	19.5 (A)	27.2 (B)	21.6 (C)
5	14.6 (E)	31.7 (A)	16.7 (B)	26.3 (C)	23.8 (D)

A conventional ANOVA F test results in a p-value of 0.05 right on the border of the commonly used significance level. However Figure 2 indicates the value 19.2 at intersection 2 and time period 3 might be an outlier and so this p-value is possibly in error. The RTF test results in a p-value of 0.03, the ARFT 0.07 and the KW 0.80. If, as in section 3, we have decided to use the RTF test, then we would decide there were significant differences. Use of the ARFT and KW tests would suggest no sequence differences.

Figure 2. Traffic Light Data



As before, the distribution of AKW is not always well approximated by χ^2. However if we are prepared to calculate a Monte Carlo p-value, here we find 0.01 for the AKW test. In this case the AKW test is most sensitive of the tests considered.

5. Replicating Latin Squares

For Latin squares of size 3 × 3 or 4 × 4, the degrees of freedom for the error term in the ANOVA are unacceptably small and so such Latin squares are often replicated. A model for replicated Latin squares is

$$Y_{ijm(k)} = \mu + \alpha_k + \beta_{i:m} + \gamma_{j:m} + \delta_m + (\alpha\delta)_{km} + E_{ijm(k)}$$

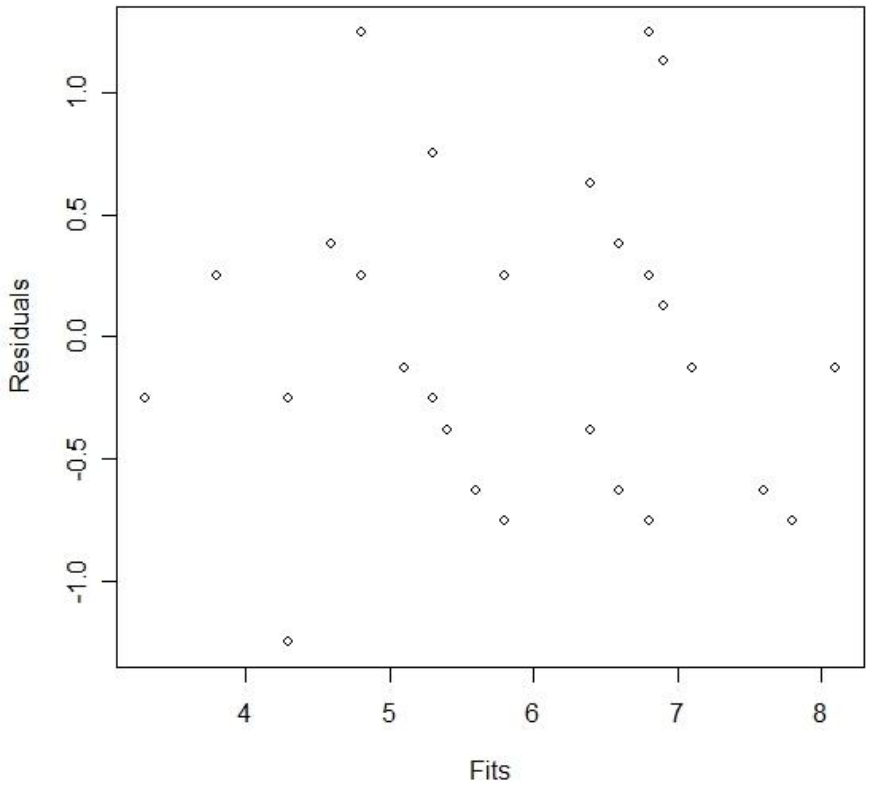
in which μ and α_k were defined above, $Y_{ijm(k)}$ indicates product k is in cell (i, j) of the mth Latin square, $E_{ijm(k)}$ is the corresponding error, δ_m is an effect due to the mth Latin square, and $(\alpha\delta)_{km}$ is the effect of the interaction between the kth product and the mth Latin square. The notation i:m and j:m denotes row and column effects defined within the mth Latin square.

Table 6. Tenderness data.

	Square 1				Square 2			
Columns	1	2	3	4	5	6	7	8
Rows								
1	D: 7	A: 7	C: 7	B: 7	A: 6	D: 7	C: 8	B: 5
2	B: 5	C: 6	A: 5	D: 7	B: 3	A: 4	D: 5	C: 4
3	A: 5	B: 7	D: 7	C: 6	D: 7	C: 6	B: 5	A: 6
4	C: 8	D: 8	B: 6	A: 5	C: 5	B: 6	A: 3	D: 6

Gacula et al. (2009, p.133) give the tenderness scores for pork loins tenderized by four different methods A, B, C and D. There are two Latin squares involved. The data are given in Table 6. Four animals make up the columns in each square and the two left loins and two right loins make up the rows.

Figure 3. Taste Test Data



As the data consists of the integers 3, 4, 5, 6, 7, 8, and as Figure 3 shows three possible outliers when the ANOVA residuals are plotted against the ANOVA fitted values, $\hat{y}_{ijm(k)}$ say, we might consider a nonparametric analysis more appropriate than the usual parametric ANOVA analysis. We might also use a nonparametric analysis if the data were originally ordered categories to which arbitrary scores were given. In any case ranking seems sensible here.

If we calculate the *ARTF* statistic a value of 5.19 is obtained compared to the ANOVA $F = 3.96$; the corresponding p-values are 0.036 and 0.016 respectively. We use the ARTF test as there is an interaction term in the model $Y_{ijm(k)}$ above. It appears the ARTF test is a little more sensitive than the ANOVA F test when the F approximation is used. To calculate the ARTF test statistic we use the aligned values $y^{*}_{ijm(k)}$ rather than the y_{ijm}, the raw data, where

$$y^{*}_{ijm(k)} = y_{ijm(k)} - \hat{\mu} - \hat{\beta}_{i:m} - \hat{\gamma}_{i:m} - \hat{\delta}_{m} - (\hat{\alpha}\delta)_{km}$$

in which

$$\hat{\mu} = \bar{y}_{...(.)},\ \hat{\beta}_{i:m} = \bar{y}_{i.m(.)} - \bar{y}_{..m(.)},\ \hat{\gamma}_{j:m} = \bar{y}_{.jm(.)} - \bar{y}_{..m(.)},$$
$$\hat{\delta}_{m} = \bar{y}_{..m(.)},\ (\hat{\alpha}\delta)_{km} = y_{..m(k)} - \bar{y}_{...(k)} - \bar{y}_{..m(.)} + \bar{y}_{...(.)}.$$

In future work it would be interesting to check whether or not the F distribution approximation for the ARTF statistic works as well as it did in the single square case. If not, Monte Carlo methods will be needed to get p-values for the between products effect. This is less convenient than using the F distribution. A check will also need to be made on whether or not the F approximation to the distribution of the ARTF statistic results in lower power when there are outliers as was indicated in Table 4 for the single Latin square. For the present data set the Monte Carlo p-value for the ARTF statistic is 0.057 as opposed to 0.016 obtained using an $F_{3,12}$ distribution; that is, at the traditional 0.05 level the F test is significant and the Monte Carlo test is not. Perhaps the difference in p-values is due to the outliers, as we have just discussed.

6. Conclusion

The rank transform method has good test sizes and powers for the 5×5 Latin square for the parameter values and error distributions employed in section 3. Satisfactory p-values can be obtained using the F distribution. As expected the Kruskal-Wallis test has poor power unless alignment is used but Monte Carlo methods are needed to obtain satisfactory p-values. Even when there are normal errors the rank transform method appears to have good power. Perhaps taking ranks avoids ‘noise’ in the raw data and helps find real differences in the products being compared. We have not looked at the performance of a permutation test here but

perhaps like the ANOVA F test such a test might be influenced by 'noise' in the raw data.

An example of replicated Latin squares is given where alignment is used prior to application of the rank transform. The alignment is meant to adjust the rank transform when there is interaction present.

References

Box, G.E.P., Hunter, J.S. and Hunter, W.G. (2005). *Statistics for Experimenters*. 2nd Edition. New York: Wiley.

Gacula, M.C., Singh, J., Bi, J. and Altan, S. (2009). *Statistical Methods in Food and Consumer Research*. 2nd Edition. New York: Academic Press.

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

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

Kuehl, R. (2000). *Design of Experiments: Statistical Principals of Research Design and Analysis*. Belmont, California: Duxbury Press.

Rayner, J.C.W. and Best, D.J. (2011). Nonparametric Tests for Two Factor Designs with an Application to Latin Squares. *Proceedings of the Fourth Annual Applied Statistics Education and Research Collaboration (ASEARC) Research Conference, February 17—18, 2011*: Parramatta, Australia.