Non-Parametric Testing: A Case Study of Extended Analysis for Ranked Data with Ties

A.A. Akomolafe*[1]; F.B. Adebola[1]; and O.I. Oseghale[2]

[1]Department of Statistics, Federal University of Technology, PMB, Akure, Nigeria.
[3]Department of Mathematics and Statistics, Joseph Ayo Babalola University, Ikeji Arakeji, Osun State, Nigeria.

E-mail: akomolafe01@yahoo.com *
femi_adebola@yahoo.com

ABSTRACT

In using non-parametric analytical tools, ranking data, which result from m raters ranking n items, are difficult to visualize due to their discrete algebraic structure, and the computational difficulties associated with them when n is large. This problem becomes worse when raters provide tied rankings or not all items are ranked. Standard analysis for ranks from two-way layout data with ties or for "rank transformed" data with ties can be extended to allow better research products comparison as well as not concealing the richness and aberration of the data. In addition to detecting product average ranks effects, the research analysis allows detection of significant nonlinear effects, umbrella effect, linear contrast and difference in distributions.

The analysis carried out was based on the soft drinks products and the results are presented for comparison of the products. There are no differences in the results obtained from the standard Friedman analysis. However, there is significant nonlinear effect which gives the manufacturer an important information towards meeting his expectation as well as the demand of the consumer which will eventually aid planning and execution by the two parties involved.

(Keywords: rank effect, linear constant, Friedman analysis, nonlinear effect, umbrella effect)

INTRODUCTION

Consumer packaged-good products, as they appear on our supermarket shelves, would have gone through a market research evaluation before the producers introduced the products fully into the market for human consumption or use. One of the market research techniques commonly use a preference ranking to compare typically, two or more products with the market leader there in this research, we will be considering comparison of soft drinks (7up, Teem, Pepsi).

Non-parametric statistical tools as it been applied in statistics to solve some specific parameter for real life problem are commonly used to verify that apparent differences between products are due to more than chance or random effects. Best (1993), in his research, proposed a new partition of a statistics been introduced by Anderson (1959). The statistics applied to ranks or rank transform data and combine with the partitioning thereby given more room to the comparison.

The extended analysis given in Best (1993) and ordered product analysis of Rayner and Best (2001) are widely applicable to market research, sensory evaluation in clinical and electoral matters and many other applications that use a randomized block design. This research widens the application area even further by considering analysis for tied ranks data. Our objective in this research his to present and explain a more dependable, reliable, and unique method in non-parametric statistics and showcasing as it is been used while avoiding technical details and proofs.

Such details have been given in Brockhoff et al. (2003). A ranking can be complete, which means all n items are ranked, or incomplete, which means some items are not ranked. A ranking, whether it is complete or incomplete, can be with-ties or without-ties; it is with-ties if some of the ranked items are not clearly preferred to others. Raters are often people but can also be computer programs; one example is search engines that provide a partial ordering of web sites. For example, suppose m members of a professional society vote for the top k of n candidates for the society council, where k < n, and supply ranks of

1 to k for their votes. Each rater establishes an ordering on all items in which there are $n-k$ ties for last place. As a result, the rankings are complete and with-ties. If $n = k$ then the rankings are complete and without-ties. If a rater can add names to the list, not all raters have the same write-in process, and the items are the original list plus write-ins, then the rankings are incomplete and with-ties.

Data obtained from consumer packaged-goods as market research data for the preference ranking is presented in the next section as well as recalling the orthogonal polynomials and associated effects used by Best (1993) for untied ranking. Linear and quadratic effects are so named because they depend on linear and quadratic polynomials. The sum of squares of the linear effects is just Friedman's classic ranking test statistics for randomized block designs to allow tied rankings when tied ranking are not allowed and when the products have been compared have an ***a priori*** ordering attached to them.

In a market research where the richness of statistical tools is pronounced, investigation came out in December 2015 in Ibadan, Nigeria, 55 consumers were asked to rank in order of preference, three soft drinks (7up, Teem, and Pepsi) products tagged A, B, and C, respectively. Adult customers who regularly opt for the products were interviewed and the order of tasting was randomized for all interviewed customers.

Tied ranking was allowed and the results are shown in Table 1. For each of customer, we assigned tied ranks by recording 1/m ranking involved in an m-ways tie. Thus, for example, customer one gives B and C half a rank of 2 and half a rank of 3 while A is given a rank of 1. Table 2 shows the soft drink products by rank matrix formed by summing overall customers.

Table 2 is not the usual type of contingency table as the sum of the ranks given by a customer is always $P\ (P+1)/_2$ where P is the number of the products. Further, all row and column sums in Table 2 equals b, where b is the number or consumers, thus Table 2 is not independent and it is not appropriate to calculate the usual chi-square statistics and to assume this has a chi-square statistics distribution with $(P-1)^2$ degree of freedom $X^2(P-1)^2$. We can, however, calculate the statistics defined in the following section, which do have convenient chi-square distribution, at least approximately. If there are no ties, the statistics we are about to define are equal to well-known statistics and we will note instances of this as they occur.

Table 1: Tied Ranking.

A	B	C	A	B	C	A	B	C
1	2.5	2.5	1	2.5	2.5	2	3	1
2	3	1	1.5	1.5	3	2.5	1	2.5
2	3	1	2	3	1	1.5	1.5	3
2.5	1	2.5	2.5	2.5	1	2.5	2.5	1
2	2	2	1.5	1.5	3	1.5	1.5	3
1	3	2	1.5	1.5	3	3	1.5	1.5
1.5	1.5	3	2	3	1	2.5	1	2.5
1.5	3	1.5	3	1.5	1.5	3	1.5	1.5
2.5	1	2.5	3	1.5	1.5	2.5	1	2.5
2	3	1	1.5	3	1.5	3	1.5	1.5
1.5	1.5	3	2	1	3	2.5	1	2.5
2.5	1	2.5	2	3	1	2	3	1
2	2	2	2.5	1	2.5	2	3	1
1.5	1.5	3	2.5	2.5	1	3	1.5	1.5
2	3	1	1.5	3	1.5	2.5	2.5	1
3	2	1	1.5	3	1.5	3	1.5	1.5
2	3	1	1.5	3	1.5	3	2	1
1	2	3	1	2	3	3	2	1
1	2	3						

Table 2: Products by Ranks Matrix.

Product	1	2	3
A	$14^{1/6}$ $16^{1/6}$	$24^{2/3}$	
B	$17^{2/3}$ $19^{1/6}$	$18^{1/6}$	
C	$23^{1/6}$ $19^{2/3}$	$12^{1/6}$	

SOME DEFINITIONS

Define, for rank j= 1,……,t and t> 1:

$$g_1(j) = \sqrt{\left(\frac{12}{t^2-1}\right)} \quad \left(j - \frac{t+1}{2}\right)$$

and, for rank j=1,….., and t >2,

$$g_2(j) = \sqrt{\left\{\frac{180}{(t^2-1)\,(t^2-4)}\right\}} \quad \left\{\left(j - \frac{(t-1)^2}{2}\right) - \frac{t^2}{12} - 1\right\}$$

These g_r (j) are polynomials orthonormal on j =n1, …., t if equal weights are used. The linear or mean effect, unadjusted for ties, for ith of t products, M_i say, may be defined as:

$$M_i = \sqrt{\left(\frac{t-1}{bt}\right)} \; \sum_{j=1}^{t} N_{ij} \;\; g_2(j),$$

And similarly the quadratic or variance effect, unadjusted for ties, for the ith of t products may be defined as:

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

Where:

N_{ij} *is the number of times product i receives rank j*

If t= 2 only the M_i values can be calculated; if t=3 both M_i and V_i values can be calculated. Often in sensory evaluation, fatigue occurs if more than four products are compared. However, if t> 3 higher order effects can still be defined in terms of higher order polynomials.

Formulae for these are given in Rayner and Best (2001). The 'unadjusted for ties' effects just defined were introduced by Best (1993). We now adjust these effects to allow for ties.

LINEAR EFFECTS ANALYSIS OF THE SOFT DRINKS EXAMPLE

For these soft drink data, we easily find that for j= 1, 2, and 3, g_1 (j) = 1.2247 (j-2). When the data contain ties, as here, the linear effects as defined in the previous section need division by an adjustment factor $\sqrt{a_i}$ say, which depends on g_1 (j) and the ties structure. Specifically,

$a_1 = g_1^T U g_1 / (bt)$,

where $= g_1^T$ (g_1 (1), g_1 (2), ….g_1 (t) and the $(I,j)^{th}$ element of U counts the number of times rank I and rank j are tied.

Suppose, for any given consumer, the ranks j_1, j_2,…….j_m tie, making up an m-way tie where $1 \leq m \leq t$. In this case 1/m is added to the elements (j_i, j_k) of the matrix U for all I, k =1,…m. This adds a total of m*m* (1/m) =m to the sum of the elements of U. An untied rank j is regarded as a one –way tie and one is added to the element (j,j) of U.

Clearly, the matrix, formed by summing over all consumers, is symmetric. For the soft drink data Table # gives U. Consumer 1, for example, contributes to U with 0.5 for the $(2,2)^{th}$, $(3,2)^{th}$ cells, and 1 to the $(1,1)^{th}$ cell. If there are no ties U= Bi, and a_1 = 1, where I_t is the t by t identity matrix.

Table 3: Rank by Ranks matrix

Product	1	2	3
1	$44^{1/6}$	$10^{1/6}$	2/3
2	$10^{1/6}$	$37^{1/3}$	$7^{1/6}$
3	2/3	$7^{1/6}$	$47^{1/6}$

Using Table 3 counts we find the adjusted effects, $M_i / \sqrt{a_i}$ for i= 1,….t, shown in Table 4.

Brockhoff et al. (2003) show that the sum of squares of the adjusted linear effects id Friedman's rank test adjusted for ties. If the response is binary, the adjusted linear effects sum of squares is Cochran's Q. For te Table 1 data, Friedman's statistic adjusted for ties 0.411 with a non-significant p-value of 0.81.

Multiple comparisons can be given using the fact that for $i \neq j, \frac{M_i - M_j}{\sqrt{a_i}}$ is approximately N (0,2). Brockhoff et al. (2003) give a proof of this. Multiple comparisons are important in deciding which is the best product to bring to market.

Table 4: Linear Adjusted Effects.

Product (i)	Linear ($M_i / \sqrt{a_i}$)
A(1)	0.2985
B(2)	0.2233
C(3)	-0.5219

It appears from the comparison of mean ranks analysis done so far, that soft drinks are not significantly different in preference and so the manufacturer may decide to market the product that have the cheapest production capacity. We now describe this analysis.

ANDERSON'S STATISTIC FOR TIED RANKS

The statistics given so far assess linear effects when there are tied ranks. It is also useful to have an omnibus statistic which potentially detects any sort of differences in the response distributions of the products being compared. For untied ranks the Anderson (1959) statistic is such an omnibus homogeneity statistic. Following Brockhoff et al. (2003) we now give a generalization of Anderson's (1959) statistic that allow for tied ranks. Let z_i be the vector of x=t-1 elements given by:

$$z_i = ((N_{i1} - y) / \sqrt{y}, (N_{i2} - y) \sqrt{y}, \ldots, (N_{ix} - y) / \sqrt{y})^T$$

where y = b/t for i= 1,…..t. The elements of z_i are of the form (observed-expected) $/\sqrt{expected,}$ as in the classic chi-squared statistic. Further, take R* + (U/b-J_t / t) where U is the matrix counts defined above and J_t is a t by t matrix with every element equal to unity. If the data are untied U/b is the identity matrix. The statistic A, which reduces to Anderson's statistic when there are no ties is given by:

$$A = \left(\frac{t-1}{t}\right) \sum_{j=1}^{t} z_i^T R^{-1} z_i$$

Where R is R* with the last row and column deleted.

Brochoff et al. (2003) also shows that if all possible pairs of consecutive ranks are untied at least once, A has an asymptotic chi-squared distribution with $(t-1)^2$ degrees of freedom: $X^2_{(t-1)^2}$. This ties condition should be met for many real data sets.

A simple way to meet the ties condition is to 'untie' one consumer's ranks by randomly assigning the tied ranks. Alternatively, a rank reduced version of the statistic can be used; see Brochoff et al. (2003) for details. For the soft drinks data we have:

$$Z_1 = (-0.9731, 1.4792)^T, z_2 = (-0.1557, -0.0389)^T, z_3 = (1.1288, -1.4402)^T$$

$$R = \begin{bmatrix} 0.4697 & -0.1485 \\ -0.1485 & 0.3515 \end{bmatrix} \text{ and } R^{-1} = \begin{bmatrix} 2.4572 & 1.0379 \\ 1.0379 & 3.2833 \end{bmatrix}$$

We find A = 8.777 and as all three pairs of ranks are untied at least once (see, for example, the second consumer), this has an approximate chi-squared distribution with four degrees of freedom. $X^2_{4.}$ Thus, the approximate p-value is 0.07. since the linear effects were found to be non-significant in the previous section, this suggests the almost significant p-value is due to nonlinear rather than linear effects.

Reference to Table 2 suggests the nonlinear effect is due to segmentation of consumer's opinions for product C or that product C was of uniform quality. Either of these conclusions is important to the food manufacturer or supplier. For tied data, we need a new second order polynomial if we wish to have a dispersion test- see the Appendix.

The calculations outlined in the Appendix may not always be needed. A fairly complete analysis may involve calculation of just A. Friedman's statistic adjusted for ties and the difference of these two statistics. This difference statistic has an asymptotic chi-squared distribution with (t-1) (t-2) degrees of freedom, $X^2_{(t-1)\ (t-2)}$, and indicate whether there are nonlinear effects which in a sensory evaluation application could be caused by market segmentation or non-uniform product. In our example, this difference has an asymptotic p-value of 0.02.

Brockhoff et al. (2003) give brief simulation studies indicating the chi-squared approximations are adequate, but we suggest that permutation tests could be done as a check for small data sets.

PAGE AND UMBRELLA TESTS

Rayner and Best (2001), for example, define Page and umbrella tests for ranked data without ties from randomized block designs. These tests assume an ***a priori*** ordering of the products. This ordering may for example, be done on the basis of historical sales figures. In an analogous manner when there are tied ranks we may define the test statistics:

$$W_{11} = \sqrt{\left(\frac{t-1}{ba_1}\right)} \sum_{i=1}^{t}\sum_{j=1}^{t} N_{ijg1}\,(i)\; g_1\,(j)\; /t$$

and

$$W_{21} = \sqrt{\left(\frac{t-1}{ba_1}\right)} \sum_{i=1}^{t}\sum_{j=1}^{t} N_{ijg2}\,(i)\; g_1\,(j)\; /t$$

for the page and umbrella tests respectively. Notice that W11 involves a linear polynomial g1(i) multiplied by another linear polynomial g1(j), and so we could say the Page statistic is a linear statistic. Similarly, the umbrella statistic can be said to be a quadratic by linear statistic. For the data in Table 1, assuming now that there is an a priori ordering of products, A > B> C, we find a Page statistic value of -0.710 and an umbrella test statistic of -0.335, with asymptotic one-tailed p-values of 0.24 and 0.37, respectively. These non-significant values agree with inspection of Table 2. The asymptotic p-values we quote use the standard normal distribution.

Notice that W11 differs from Page statistic produced by the IMSL (1995) routine FRDMN or by StatXact (1995). Neither IMSL nor StatXact really adjusts for ties when calculating p-values based on the chi-squared approximation, and so we prefer to use the Page statistic adjusted for ties we have just given.

To emphasis that the method of adjusting for ties is important, consider the data in Table 5 for these data IMSL gives a Page statistic p-value of 0.063. StatXact gives an asymptotic p-value of 0.047 and an exact value of 0.037, while using our ties adjusted Page statistic we find the asymptotic p-value is 0.031. The first of these p-values might not be considered significant and all three asymptotic p-values disagree.

Table 5: A Second Set of Product Rankings.

Consumer	Product A	B	C
1	1.5	1.5	3
2	1	3	2
3	1.5	1.5	3
4	1.5	1.5	3
5	2	3	1
6	1.5	1.5	3
7	2.5	1	2.5
8	2.5	1	2.5
9	1	2.5	2.5
10	2	1	2
11	2.5	1	2.5
12	2	2	2

PARTITIONING FRIEDMAN'S STATISTIC

Suppose we let ***S*** be Friedman statistic calculated for tied data, namely:

$S=(\sum_{i=1}^{t} M_i^2)/a_1.$

If we want to know, for the Table 1 data, whether the average rank for product C differs from the average of the average ranks for products A and B we can form linear contrast, L say, given by:

$1.5a_1L = \{M_3 - (M_1 + M_2)/2\}^2.$

For Table 1 data L = 0.408 and S – L = 0.003 with p-values, based on the approximating X^2_1 distribution, of 0.523 and 0.956, respectively. The constant, 1.5, is derived, as is usual with linear contrast, in this case $1.5 = 1^2+2(0.5)^2$. There is no area that C's average is that of the A and B ranks. If the data are binary, then *L* is a Miettinen-type statistic. Linear contrasts could also be used to partition the dispersion statistic, or other linear contrasts, as appropriate, could be examined.

CONCLUSION

In this research, we have considered randomized block or two-way layout ranked data and proposed new statistics for testing (a) differences in distributions of ranks with ties, (b) differences in nonlinear effects of ranks with ties, (c) whether an 'umbrella' ordering exists when there are tied ranks and (d) linear contrasts. We gave a sensory evaluation example where the data were ranks. In some applications the data may be obtained as category rating data or as continuous line scale data and it may be appropriate to rank such data. This might be because, for example, the categories are not equispace and so assigning scores 1,2,3…. are not valid. Alternatively, a consumer who gives continuous line scale scores 20 and 40 may not really mean that one product had twice the flavor of the other. Using ranks also ensures that a consumer who gives scores 40 and 60 has the same impact as a consumer who spreads these to 10 t0 90.

APPENDIX

If we require a dispersion test and the data are tied then we need to redefined g_2 =b (g_2 (j)). Initially take $g_2(j) = j^2 + c_1 j + c_0$ for j= 1,….t, where the constants are to be determined. The orthogonality require $\sum_{j=1}^{t} g_2(j) = 0$ and $g_1^T U g_2 = 0$. This allows us to solve two linear equations for the two unknown constants. We also require a normalizing constant, say, E, such that with $g_2(j) = E^{0.5}(j^2 + cij + c_0)$, $g_2^T (U/b)\, g_2 = t.$

Similarly, if t> 3, we can redefined g_3, g_4, and so on. If the redefined g_r = (g_r (j)) are used to defined the order r effects statistic then the adjustment factors discussed earlier are not needed. Brockhoff et al. (2003) discusses the approach in this Appendix in greater detail.

REFERENCES

1. Anderson, R.L 1959. "Use of Contingency Tables in the Analysis of Consumer Preference Studies". *Biometrics*.15:582-590.

2. Best, D.J. 1993. "Extended Analysis for Ranked Data". *Aust. J. Statist*. 35:257-262.

3. Brockhoff, P.B., D.J. Best, and J.C.W. Rayner. 2003. "Anderson's Test with Tied Ranks". *J. Statist. Planning Infer*.

4. Critchlow. D.E. 1985. "Metric Methods for Analyzing Partially Ranked Data". *Lecture Notes in Statistics*. 34, Springer.

5. David, H. 1988. *The Method of Paired Comparisons.* Oxford Press: London, UK.

6. Diaconis, P. 1988. *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics: Bethesda, MD.

7. Diaconis, P. 1989. "A Generalization of Spectral Analysis with Application to Ranked Data". *Annals of Statistics*. 17(3):949–979.

8. Fligner, M.A. and J. S. Verducci, editors. 1993. *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag: Berlin, Germany.

9. Gabriel, K. 1982. "Biplot". *Encyclopedia of Statistical Science*. 1:263–271.

10. Goldberg, K., T. Roeder, D. Gupta, and C. Perkins. 2001. "Eigentaste: A Constant Time Collaborative Filtering Algorithm". *Information Retrieval*. 4(2):133–151.

11. IMSL. 1995. *IMSL Fotran Numerical Libraries*. Visual Numerics: Houston, TX.

12. Inselberg, A. 1989. "Visualizing Multi-Dimensional Structure using Parallel Coordinates". *American Statistical Association Proceedings of the Section on Statistical Graphics*.

13. Kendall, M.G. 1995. "A New Measure of Rank Correlation". *Biometrika*. 30(1/2): 81-93

14. Rayner, J.C.W. and D.J. Best. 2001. *A Contingency Table Approach to Nonparametric Testing*. Chapman and Hall/CRC: Boca Raton, FL.

15. Cox, T.F. and M.A. Cox. 1994. *Multidimensional Scaling*. CRC Press: Boca Raton, FL.

SUGGESTED CITATION

Akomolafe, A.A., F.B. Adebola, and O.I. Oseghale. 2017. "Non-Parametric Testing: A Case Study of Extended Analysis for Ranked Data with Ties". *Pacific Journal of Science and Technology*. 18(2):239-245.

Pacific Journal of Science and Technology