TESTS FOR INTERACTION IN A TWO-WAY LAYOUT: SHOULD THEY BE INCLUDED IN A NONPARAMETRICS COURSE?

Thomas P. Hettmansperger and Ryan Elmore
Penn State University
USA

We first note that tests for interaction are missing in virtually all textbooks on nonparametric statistics. We will discuss some reasons why this is so. We then make a case for featuring tests for interaction in the course. By learning how to use median polish and graphical displays students can begin to conceptualize what an interaction means. This will strengthen their understanding of additive models as well. After a conceptual basis for understanding interaction is in place, we can then proceed to design tests for interaction. They will not be strictly nonparametric. This provides a good opportunity for discussion of what it means to have a nonparametric test and why it is impossible to construct an ordinary permutation test for interaction.

INTRODUCTION

The purpose of this paper is to discuss a rank test for interaction in a two-way layout with m observations per cell. The test can be incorporated into a course in applied nonparametric statistics given as a second course in statistics at American universities. A quick review of recent books in applied nonparametric statistics did not turn up a single case in which such a test is presented. See for example, Hollander and Wolfe (1999), Lehmann (1998), Conover (1980) and Sprent and Smeeton (2001). Since a strictly distribution-free test for interaction in the traditional parametric model is not possible, published work in journals has concentrated on either redefining the concept of interaction or redefining the model. See Marden and Muyot (1995) and Brunner and Puri (1996), respectively. Here we wish to stick to the accepted parametric model and discuss a natural test that, surprisingly, has not been explicitly presented in the literature. A rank transform test in which all the data are ranked together has been suggested by Conover and Iman (1981); however, this has been shown to give misleading results by McKean and Vidmar (1994) among others and is not recommended.

A RANK TEST FOR INTERACTION

We begin with the traditional model: $Y_{ijk}= \mu + \alpha_i + \beta j + \gamma_{ij} +\varepsilon_{ijk}$ with i= 1,…,b, j=1,…c, and k=1,…,m. We assume the errors are independent and identically distributed from some continuous distribution with median 0. Further, $\Sigma\alpha_i = \Sigma\beta_j = \Sigma_i\gamma_{ij} = \Sigma_j\gamma_{ij} = 0$. In a traditional analysis of variance we decompose the total sum of squares:

$$\sum\sum\sum(Y_{ijk} - \overline{Y}_{...})^2 = \sum\sum\sum(Y_{ijk} - \overline{Y}_{ij.})^2 + m\sum\sum(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2$$
$$+ mc\sum(\overline{Y}_{i..} - \overline{Y}_{...})^2 + mb\sum(\overline{Y}_{.j.} - \overline{Y}_{...})^2 \quad (1)$$

and the F test statistic for testing H_0: $\gamma_{ij} = 0$ for all i,j is defined by:

$$(b-1)(c-1)F = m\sum\sum(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2 / s_\varepsilon{}^2 \quad (2)$$

where $s_\varepsilon{}^2$ = MSE = $\sum\sum\sum(Y_{ijk} - \overline{Y}_{ij.})^2$ /(bc(m-1). When the error distribution is Gaussian, F has an F distribution with (b-1)(c-1) and bc(m-1) degrees of freedom. Otherwise, (b-1)(c-1)F is approximately distributed as a chisquare random variable with (b-1)(c-1) under mild regularity conditions.

We note two aspects of the test statistic: (a) The numerator is a sum of squares of cell mean residuals formed by removing estimates of the row, column, and overall effects and (b) The denominator is an estimate of the error variance and acts to scale the data. We approach the rank test statistic in a similar way. We first form residuals by subtracting estimates of the main effects from the observations and then we rank the residuals. Let $R_{ijk} = Rank(Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)$ where

the estimates are not necessarily least squares estimates. We will discuss alternative estimates below. Note also that $\bar{R}_{...} = (N+1)/2$ where N = mbc. Next, we analyze the variance of ranks:

$$\sum\sum\sum(R_{ijk} - \frac{N+1}{2})^2 = \sum\sum\sum(R_{ijk} - \bar{R}_{ij.})^2 + m\sum\sum(\bar{R}_{ij.} - \bar{R}_{i..} - \bar{R}_{.j.} + \frac{N+1}{2})^2$$
$$+ mc\sum(\bar{R}_{i..} - \frac{N+1}{2})^2 + mb\sum(\bar{R}_{.j.} - \frac{N+1}{2})^2. \quad (3)$$

Then the rank test statistic for interaction is similar to (2):

$$Q = m\sum\sum[\bar{R}_{ij.} - \bar{R}_{i..} - \bar{R}_{.j.} + (N+1)/2]^2 / s_R^2 \quad (4)$$

where now s_R^2 = mcb(mcb+1)/12, since the left side of (3) is simply the centered sum of squares of the integers from 1 to mbc. For large m, Q is approximately distributed as a chisquare with (b1)(c-1) degrees of freedom. The theory is detailed in Hettmansperger and McKean (1998). Further, the approximation works well for moderate to small m. The statistic Q tests to see if the cell rank averages are additive, similar to the F test. Under the null hypothesis of no interaction, they should be roughly additive; since, the estimates of main effects have been removed.

The test based on Q is not an exact randomization or permutation test in either the randomized block design or a two factor factorial experiment. This is because the column and/or the row mean ranks are not invariant under permutations. However, if the permutations are carried out and sampled in order to approximate a p-value, the results are quite close to the asymptotic approximation mentioned above. See Davison and Hinkley (1997) for a brief discussion.

Residuals can be formed using methods other than least squares; however, a least squares fitting algorithm along with a ranking algorithm are all that is needed to implement the test based on Q. Median polish is another method of producing residuals in a two-way layout; see Emerson and Hoaglin (1983) for a thorough discussion. When this is used on the original data, the rows and columns (including the main effects) all have median 0 rather than mean 0 as in the case of least squares. Median polish is available in Mintab statistical software. Median polish is similar to fitting by least absolute deviations and is more robust than least squares. A third approach is to form the residuals based on R-estimates of the main effects. In this case the residuals are adjusted so that $\bar{R}_{i..} = \bar{R}_{.j.} = (N+1)/2$. This can be achieved using Minitab and casting the two-way layout as a regression model. The command in Minitab is RREG. When R-estimates of main effects are used Q reduces to a Kruskal-Wallis statistic computed on the bc cells of the layout:

$$Q = \frac{12m}{bcm(mcm+1)}\sum\sum(\bar{R}_{ij.} - \frac{N+1}{2})^2. \quad (5)$$

Recall that the Kruskal-Wallis statistic is a rank statistic for testing treatment effects in a one-way layout. In the next section we illustrate these approaches on a data set.

EXAMPLE

The example is taken from a paper by Hahn, Haber, and Fuller (1973) and discussed by Scheirer, Ray, and Hare (1976) where the data can be found. The subjects consisted in 42 pairs of mice, selectively bred for brain weight (small, medium, and large) and raised in two different environmental conditions. After maturation the fighting behavior of pairs from the same brain weight and environment was observed. Measurements consisted of seconds of tail rattling per seconds of fighting. Tail rattling is a measure of aggression in mice. We might consider the environments as blocks and ask if there is a difference in aggression among the brain weight groups. However, we should first test for the presence of interaction. When interaction is present we will have to be careful how we assess differences in brain weight groups across the environments. The presence of interaction will entail a different pattern of differences across the blocks.

Prior to carrying out a formal test for interaction it is useful to see boxplots of the cell data. We further embed 85% nonparametric confidence intervals for the medians in the boxplots. If the 85% confidence intervals are disjoint we can reject (at roughly 5%) the null hypothesis of equality of medians. There is no attempt to control the overall error rate in this rough graphical mode. The boxplots will reveal significant pairwise differences and suggest interaction when it is present.

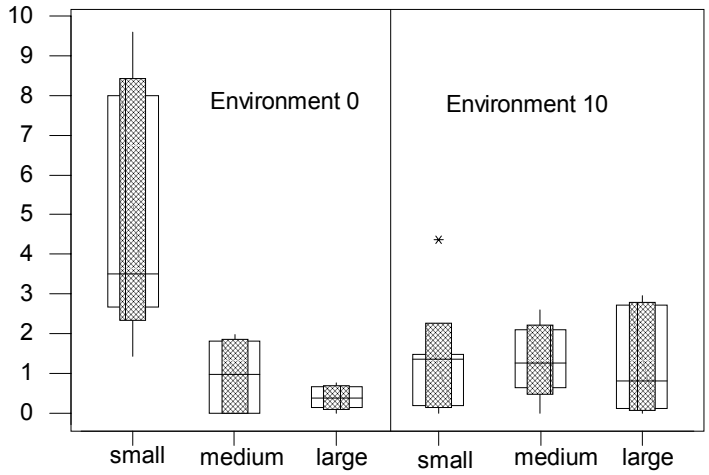

Figure 1. Boxplots of Cell Data with 85% Confidence Intervals

We see at once from Figure 1 that the confidence interval for the small brain group in environment 0 is disjoint from the other two confidence intervals. This seems to be the only source of significance. Hence, we would expect to detect interaction which would reflect the significance in the environment 0 block and no significance in the environment 10 block. We can further see a difference in variation from group to group and there is an outlier as well. We next consider a formal test for interaction.

In Table 1 we have the mean ranks of the least square residuals when the main effects have been removed. We have b = 2 environments, c = 3 brain weights, and m = 7 observations per cell. The test for interaction is highly significant and confirms the informal analysis based on the boxplots with confidence intervals. Note that the rank averages in the margins would be expected to be roughly 21.5 if the estimates of the main effects were equal to the R-estimates. As a check we approximated the permutation test. We permuted the data within blocks and recomputed the test statistic Q. This was done 100,000 times and the approximate p-value was the number of times Q was greater than or equal to 14.976 divided by 100,000. We found a value of .00044 which compares favorably to the p-value of .0008 found from the chisquare approximation.

Table 1

Average Ranks of Least Squares Residuals and Test for Interaction

	Brain Weight			
	Small	Medium	Large	All
Environment 0	26.9	17.4	15.1	19.8
Environment 10	9.6	29.9	30.1	32.2
All	18.2	23.6	22.6	21.5
Q = 14.976 p-value = .0006 (based on chisquare with 2 df)				

In Table 2 we provide the average ranks of residuals based on R-estimates of the main effects. Note that in this case Q is the Kruskal-Wallis one-way layout statistic computed on the 6 cells. The degrees of freedom are 2, not 5 associated with a one-way layout.

Table 2
Average Ranks of R-estimate Residuals and Test for Interaction

	Brain Weight			
	Small	Medium	Large	All
Environment 0	31.6	17.3	15.7	21.5
Environment 10	11.4	25.7	27.3	21.5
All	21.5	21.5	21.5	21.5
Q = 14.20 p-value = .0008 (based on chisquare with 2 df)				

Having rejected the null hypothesis of no interaction, we next consider two one-way layouts. We compute Kruskal-Wallis tests for each block separately. The null hypothesis is that there is no difference in aggression across the brain weight groups. Now ranking within each block separately, the Kruskal-Wallis statistic is:

$$KW = \frac{12m}{cm(cm+1)} \sum (\overline{R}_{j.} - \frac{N+1}{2})^2.$$

The statistic, under the null hypothesis, is approximately chisquare distributed with c-1 degrees of freedom. An exact permutation distribution can be easily determined, but the chisquare approximation is adequate for most purposes. For Environment 0, KW = 12.83 with a p-value of .002, while for Environment 10, KW = 0.16 with a p-value of .925. Hence, we have confirmed what was suggested by the CI-boxplots and the significant test for interaction. If we carry this further with multiple comparisons in each block we find that the small brain weight group is significantly different from the other two in Environment 0, as expected. Error rates can be controlled throughout by budgeting the error across the blocks and then within the Environment 0 block for the multiple comparisons.

CONCLUSIONS

We have proposed a simple rank test for interaction. The test is constructed from least squares residuals formed by removing estimates of the main effects. It is easy to implement and should be an integral part of any applied nonparametrics course. The p-value is approximated using a chisquare distribution. As an alternative, a permutation test version can be approximated, and in all examples that we have tried the p-values are close to those determined through the chisquare approximation. We generally use Minitab for computations. There is a rank regression command in Minitab that allows us to fit a general linear model based on R-estimates. Hence, any experimental design that can be cast as a regression problem can be analyzed in a similar way. S-plus functions to compute the test statistic and simulate the permutation distribution are available from the second author. Finally, there is a website where these calculations can be carried out: http://www.stat.wmich.edu/slab/RGLM/index.html.

REFERENCES

Brunner, E., & Puri, M.L. (1996). Nonparametric methods in design and analysis of experiments. In S. Ghosh and C.R. Rao, (Eds.), *Handbook of statistics 13* (pp. 631-703). Amsterdam: Elsevier.

Conover, W.J. (1980). *Practical nonparametric statistics* (2nd edn*)*. New York: John Wiley.

Conover, W.J., & Iman, R.L. (1981). Rank transform as a bridge between parametric and nonparametric statistics (with discussion). *American Statistician 35,* 124-133.

Davison, A.C., & Hinkley, D.V. (1997). *Bootstrap methods and their applications.* Cambridge: Cambridge University Press.

Emerson, J.D., & Hoaglin, D.C. (1983). Analysis of two-way tables by medians. In D.C. Hoaglin, F. Mosteller, and J.W. Tukey, (Eds.), *Understanding robust and exploratory data analysis* (pp. 166-207). New York: John Wiley & Sons.

Hahn, M.E., Haber, S.B., & Fuller, J.L. (1973). Differential anatogonistic behavior in mice selected for brain weight. *Phys. and Behavior 10,* 759-762.

Hettmansperger, T.P., & McKean, J.W. (1998). *Robust nonparametric statistical methods.* London: Arnold.

Hollander, M., & Wolfe, D.A. (1999). *Nonparametric statistical methods* (2nd edn). New York: John Wiley & Sons.

Lehmann, E.L. (1998). *Nonparametrics: Statistical methods based on ranks.* Upper Saddle River, NJ: Prentice Hall.

Marden, J.I., & Muyot, M.E.T. (1995). Rank tests for main and interaction effects in analysis of variance. *J. Amer. Statist. Assoc., 90,* 1388-1398.

McKean, J.W., & Vidmar, T.J. (1994). A comparison of two rank based methods for the analysis of linear models. *Amer. Statistician, 48,* 220-229.

MINITAB. (1994). *MINITAB Reference Manual.*

Scheirer, C.J., Ray, W.S., & Hare, N. (1976). The analysis of ranked data derived from completely randomized factorial designs. *Biometrics, 32,* 429-434.

Sprent, P., & Smeeton, N.C. (2001). *Applied nonparametric statistical methods* (3rd edn). London: Chapman & Hall/CRC.