DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 1 | Issue 1 Article 13

5-1-2002

Six Modifications Of The Aligned Rank Transform Test For Interaction

Kathleen Peterson
Macomb Intermediate School District and Oakland University, Michigan

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
Peterson, Kathleen (2002) "Six Modifications Of The Aligned Rank Transform Test For Interaction," *Journal of Modern Applied Statistical Methods*: Vol. 1 : Iss. 1 , Article 13.
DOI: 10.22237/jmasm/1020255240
Available at: http://digitalcommons.wayne.edu/jmasm/vol1/iss1/13

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

Journal Of Modern Applied Statistical Methods
Winter 2002, Vol. 1, No. 1, 100-109

Copyright © 2002 JMASM, Inc.
1538 - 9472/02/$30.00

Six Modifications Of The Aligned Rank Transform Test For Interaction

Kathleen Peterson
Macomb Intermediate School District, Michigan,
& Oakland University

Testing for interactions in multivariate experiments is an important function. Studies indicate that much data from social studies research is not normally distributed, thus violating that assumption of the ANOVA procedure. The aligned rank transformation test (ART), aligning using the means of columns and rows, has been found, in limited situations, to be robust to Type I error rates and to have greater power than the ANOVA. This study explored a variety of alignments, including the median, Winsorized trimmed means (10%) and (20%), the Huber$_{1.28}$ M-estimator, and the Harrell-Davis estimator of the median. Results are reported for Type I errors and power.

Keywords: ANOVA, Interactions, Aligned rank transform, Nuisance parameter

Introduction

Conover and Iman (1981) suggested a rank transform test (RT) that ranks the data before doing an ANOVA as a bridge between parametric and non-parametric statistics. However, the RT was found to be erratic with respect to both Type I and Type II errors as a test of interaction in the context of a 3 x 4 design (Blair, Sawilowsky, & Higgins, 1987) and a 2x2x2 design (Sawilowsky, Blair & Higgins, 1989). Sawilowsky and Blair (1987) commented: "Not only was the test dramatically non-robust at times, but it also demonstrated very poor power properties in many situations. This was particularly true under those conditions in which interactions were present." (p. 13)

In a review of existing non-parametric tests for interactions, Sawilowsky (1990) narrowed the search for the best test down to five: Bradley's Collapsed and Reduced technique (1979), adjusted (or aligned) rank transform, (Blair & Sawilowsky, 1990), Puri and Sen L (1985), Shoemaker's extended median test (1985), and the Hettmansperger test (1984). Sawilowsky commented on the computational difficulty of the Hettmansperger test, and pointed out that of the other four, the adjusted [aligned] rank transform appears to reach desirable power properties with the smallest sample size.

Kelley and Sawilowsky (1997) found good results for the adjusted rank test. Their study indicated that this test aligned by means had superior power properties when compared to the ANOVA if the distribution is heavy-tailed or skewed, and the F test has only a slight power advantage when testing for interactions if the populations are symmetric with light tails.

It has been noted that there were some minor inflations with regard to Type I errors in layouts higher than the 2x2. For example, with nominal alpha set to .05, null interactions in the presence of non-null main effects resulting in Type I error rates as high as .065. The question arises whether some other estimate of the nuisance parameter, other than the arithmetic mean, might better preserve the Type I error rate. The study described here followed suggestions by Toothaker and Newman(1994) and Sawilowsky (1990) for further study of the aligned rank transform test for interaction using alignments other than the mean.

Kathleen Peterson is a consultant for the Macomb Intermediate School District. Her duties include working with Macomb County secondary teachers who are integrating statistics into their math/science programs. She is also an adjunct faculty member at Oakland University.

Methodology

This Monte Carlo study of a 3x4 design was designed to examine the Type I error rate and power of six alignment statistics and the F statistic, when sampling from a variety of normal and non-normal distributions.

The six statistics used for alignment purposes were: the sample mean (ART_m), the sample median (ART_{md}), the lightly trimmed (2x10%) Winsorized mean (ART_{tml}), a heavily trimmed (2x20%) Winsorized mean (ART_{tmh}), the Huber $\psi_{1.28}(ART_H)$(Hoaglin, Mosteller, & Tukey, 1983), and the Harrell-Davis (1982) estimator of the median (ART_{HD}).

For the ART_m, estimates of the main effects were removed by calculating the means for each row and column of data. Then the mean for each row was subtracted from the observations in that row. After that, the mean of each column was subtracted from the remaining values in that column. After alignment the remaining values were ranked; then an ANOVA was done on the ranks to test for an interaction. Other alignments were done in a similar manner. Alpha levels of .05 and .01 were used.

In a search for a statistic to be used in situations where normality is not assured an important issue is: "What distributions should be studied?" There have been arguments for using real data sets (e.g., Stigler, 1977; Micceri, 1989; Sawilowsky & Hillman, 1992). Wilcox (1995)

argued for using theoretical distributions with salient features (such as kurtosis or skewness) motivated by theoretical considerations, and considering what happens when these features are altered. But Micceri (1986) pointed out in a study of 440 large data sets from social science research that in some cases, (although he used a variety of quantitative techniques to assess tail weights, asymmetry, and modality), classification could only be done by visual inspection of the pseudo-population (large sample) or a combination of visual inspection and quantitative assessments.

Micceri (1989) also pointed out that the data sets which exhibited extremely light tails (similar to the uniform distribution) tended to be asymmetric, suggesting that simulated studies based on such symmetric mathematical functions such as the uniform, logistic, double exponential, Cauchy, and t with few degrees of freedom may not represent real-world data to any reasonable extent.

Although there are an infinite number of non-normal distributions, having knowledge that a statistic is appropriate for many situations encountered in social studies research is more reassuring than knowing that a statistic works with some theoretical distributions, especially when sample sizes may not be large enough to determine if the population studied has those characteristics. For this reason, this study was done using, besides the Normal distribution, large, real data sets typical of those commonly found in social studies research.

A data set from Micceri's 1986 study (referred to as the Extremely Asymmetric Data Set) with n = 2,768, was used for the simulation. It was assumed that this data set and the subsequent ones listed were large enough to proxy a population.

Another data set from this study, with n = 5,375 and referred to as the Smooth Symmetric Data Set, is typical of gain scores, which usually showed some degree of symmetry but often had heavy tails.

Micceri found that 81.2% of the 440 data sets showed considerable or extreme lumpiness or digit preference. A data set from this group, with n = 467, referred to as the Multi-modal and Lumpy Data Set was also used.

Another Micceri achievement test data set used is the Discrete Mass at Zero with Gaps set, with n = 2,429. This data set is typical of data where there is a pretest in which one subgroup has not been exposed to the material tested and the other group has some familiarity with the subject.

A data set with n = 887, referred to as the Likert Scale data set, is data from a medical rehabilitation setting (Nanna & Sawilowsky, 1998). This set used a seven-point Likert scale.

Because previous studies (Sawilowsky, Blair, & Higgins, 1989) indicated that some rank transformation tests for interaction break down in the presence of main effects, the following effect conditions were studied:

a. Condition 1: all effects null.
b. Condition 2: main effects with no interaction with $b_3 = a_1 = c\sigma$ and $b_1 = b_2 = b_4 = a_3 = -c\sigma$, where c = .25-2.5(.25) and represents the shift.
c. Condition 3: no main effects and a disordinal interaction with $(ab)_{11} = (ab)_{12} = (ab)_{33} = (ab)_{34} = c\sigma$, and $(ab)_{13} = (ab)_{14} = (ab)_{31} = (ab)_{32} = -c\sigma$.
d. Condition 4: ordinal interaction with two main effects, with $(ab)_{11} = .5c\sigma$ and $a_1 = (ab)_{14} = -c\sigma$.

A Monte Carlo program was written as a *Minitab* (1998) Release 12.1 "macro", to take advantage of some existing *Minitab* routines. *Minitab* macros trims1.mtb, os.mtb, and hd.mtb from Wilcox (1996) were used. A problem arose relative to the os.mtb macro, used for the Huber statistic. When a data set has a large number of ties, especially near the center of the data set, it is possible for the MAD (the Median Absolute Deviation from the median) to be zero. The program was modified so that in these cases the median was used as the one-step estimator of the Huber $\psi_{1.28}$, because the median is the starting point for the iterative process determining the Huber$\psi_{1.28}$ (Hoaglin, Mosteller, & Tukey, 1983).

Samples sizes of 5, 10, 15 and 20 per cell were used. There were 5,000 repetitions for each experimental combination.

Results

The results of the Monte Carlo study are reported by effect condition. Condition 1 has all effects null, and Condition 2 has main effects with no interaction. Therefore the concern with these two conditions is the Type I error rates for the interaction.

Condition 1

Figure 1 displays an over-all view of the Type I error rates, by the aligning statistics, for all distributions and all sample sizes for alpha = .05. It shows a slight over-all tendency for the alignment statistics studied to inflate alpha. The F statistic, the ART_m and ART_{md} are the best, in that order, with the ART_m having only one value violating the stringent definition for robustness, $\alpha \pm .1\alpha$, based on a sample size of 5000. For the F statistic, all values meet the stringent definition of robustness. All the other statistics except the ART_{md} have some values violating the moderate ($\alpha \pm .25\alpha$) criterion for robustness, with the ART_{HD} being the worst.

For the ART_m, ART_{md}, ART_{trl} and ART_{trh} any

values beyond the stringent boundary were for the Extreme Asymmetry Data. The ART_H and the ART_{HD} had liberal rates for almost all the distributions studied. The F statistic was robust in all cases. Exact figures for each combination of sample size, statistic, distribution and alpha level are available at kpeterson@misd.net.

The results for Condition 1, Alpha = .01 were similar, although slightly more liberal. In addition to the elevated rates for the Extreme Asymmetry data, there were violations of the stringent goal for the ART_{md} and ART_{trl}, and the moderate goal for the ART_{trh} with the Likert distribution. There were liberal figures for most distributions for the ART_H and the ART_{HD}, although it should be pointed out that the worst violation was a rate of .0358, for the ART_{HD} with the Extreme Asymmetry Data Set.

Condition 2

For Condition 2 (no interaction but main effects) all the statistics, including the F statistic, displayed a slight tendency to inflate alpha, but all the F rates were within the limits for a stringent definition of robustness. The ART_{md} had only one out of 240 rates extreme enough to fall in the moderate interval (at n = 20 for the Likert distribution). Only the ART_{trl} (for the Extreme Asymmetry distribution, at n = 5) had a value beyond the limit for moderate robustness.

For Condition 2 (no interaction with two main effects), alpha .01, all seven statistics had problems with the Extreme Asymmetry data for n = 5. The F statistic had one rate (out of 240) beyond the liberal level (with the Extreme Asymmetry data); the ART_{md} had three rates which didn't meet stringent criteria (one of them was with the Likert data); the ART_{trl}, ART_{trh} and ART_H had almost all rates for the Extreme Asymmetry data too large for moderate robustness. Each statistic except the F statistic had at least one (but no more than three) violations with the Likert data set. The violations tended to lessen in number and severity as sample size increased. Again, to keep perspective, the largest value was .0198, for the ART_{trl} with the Extreme Asymmetry data set.

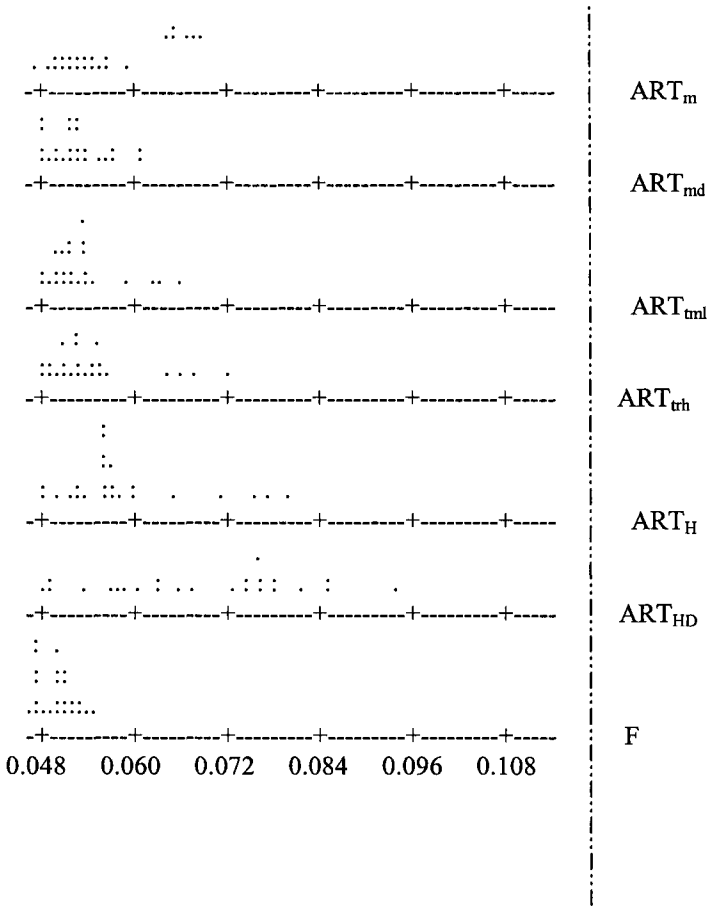

Figure 1. Type I error rates by statistic for Condition 1 (all-effects-null) with nominal alpha = .05. All four sample sizes are grouped together.

Condition 3

Condition 3 is a disordinal interaction with no main effects. Figure 2 displays histograms which show the differences in power for each statistic in comparison to the F statistic for each of the 240 sample size/ distribution/ shift level combinations for alpha = .05.

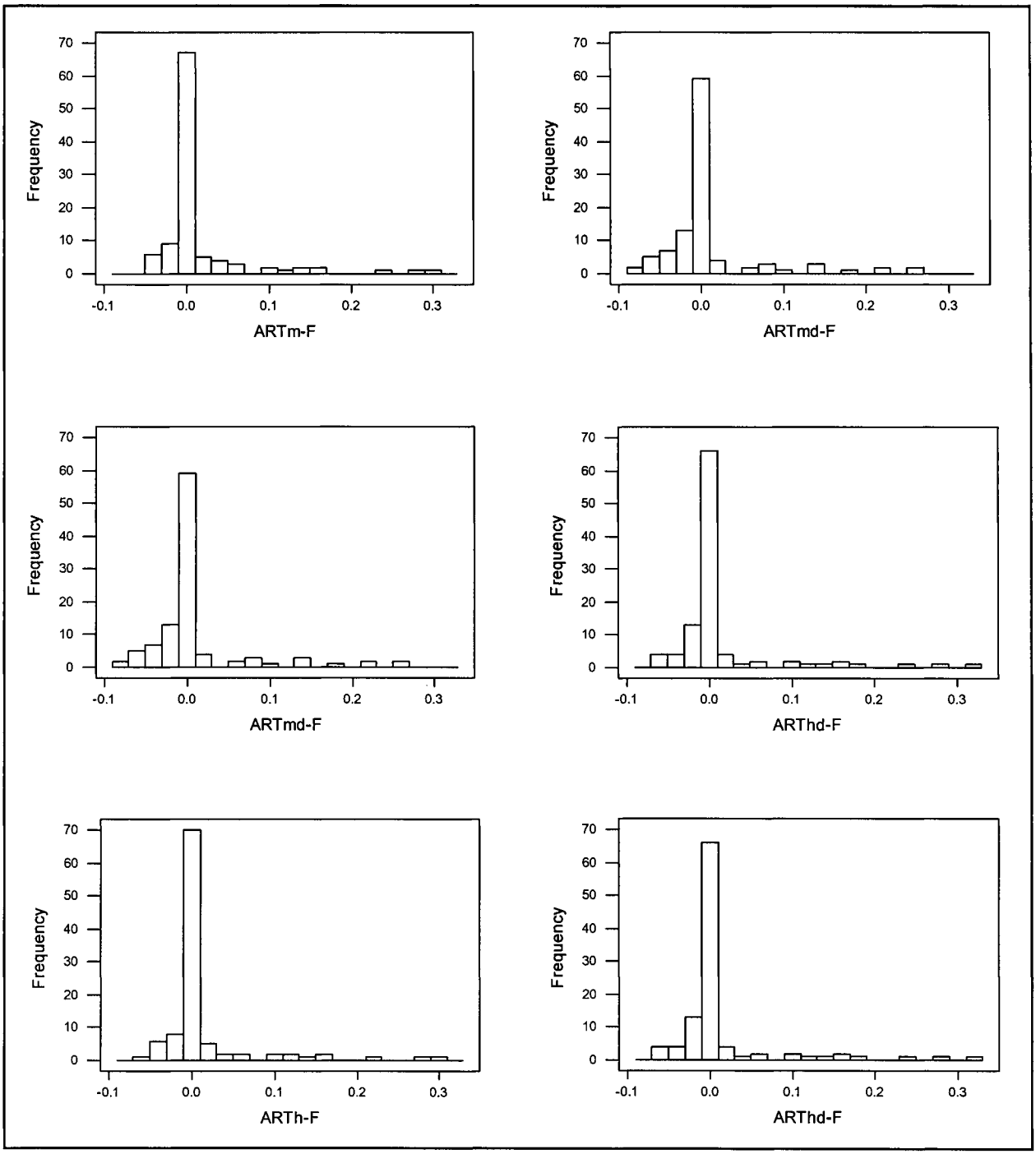

Figure 2. Histograms for Differences: Alignment Statistic Power Minus F Statistic Power (Condition 3, Alpha .05).

Examination of the histograms reveals that:

1. In most cases, differences between a given statistic and the F statistic are minimal, very close to zero.
2. For all six statistics, the data is skewed to the right, indicating that there are some cases where the statistic in question is much more powerful than the F statistic.
3. The three statistics with the heaviest and longest left tails (indicating less power than the F statistic) are the ART_{md}, ART_{H}, and the ART_{HD}.

Table 1. Descriptive Statistics for Power Differences: Alignment Statistic Minus F Statistic for Condition 3, Alpha = .05.

Variable	N	Mean	Median	StDev	Minimum	Maximum	Q1	Q3
ART_{m}-F	104	0.01664	0.0000	0.05826	-0.0408	0.3066	-0.00155	0.00600
ART_{md}-F	104	0.01047	0.0000	0.06065	-0.0840	0.2556	-0.01235	0.00015
ART_{trl}-F	104	0.01608	0.0000	0.05731	-0.0442	0.2974	-0.00120	0.00455
ART_{trh}-F	104	0.01415	0.0000	0.05313	-0.0464	0.2900	-0.00115	0.00435
ART_{H}-F	104	0.01466	0.0000	0.05773	-0.0536	0.2960	-0.00235	0.00350
ART_{HD}-F	104	0.01321	0.0000	0.06163	-0.0654	0.3114	-0.00490	0.00040

There were 48 power graphs generated for Condition 3 (6 distributions, 4 sample sizes, 2 alpha levels). Figures 3 and 4 show several of the situations where the alignment statistics show considerably more power than the F statistic. The ART_{md}, which had shown good type I error rates, showed a lack of power in some cases. This became more pronounced for Condition 4.

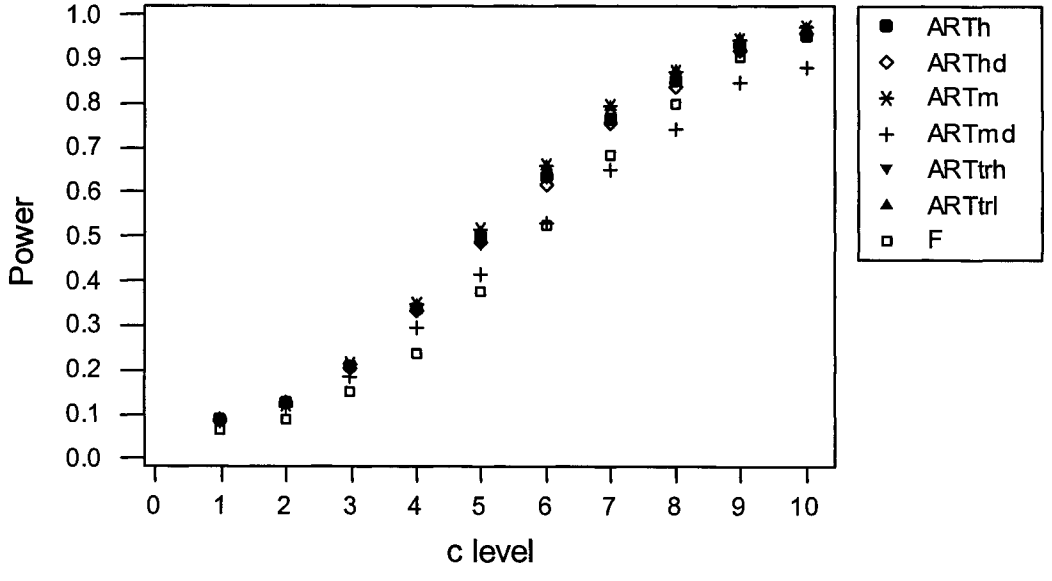

Figure 3. Power graph for Condition 3 (no main effects with a disordinal interaction) for the Extreme Asymmetry data set with nominal alpha = .05 and n = 5. The C level is the multiple of .25σ used for shift.

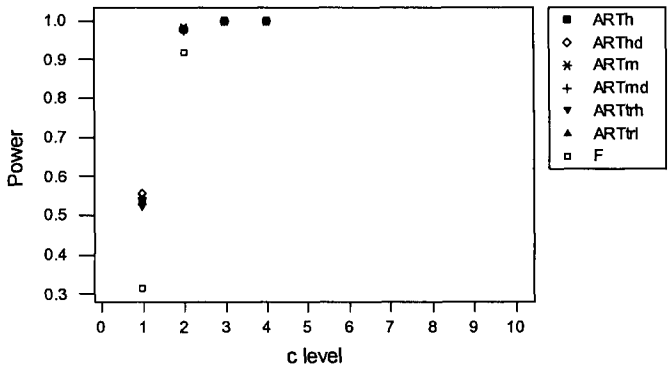

Figure 4. Power for Condition 3 (no main effects, disordinal interaction, Extreme Asymmetry), $\alpha = .05$, $n = 10$.

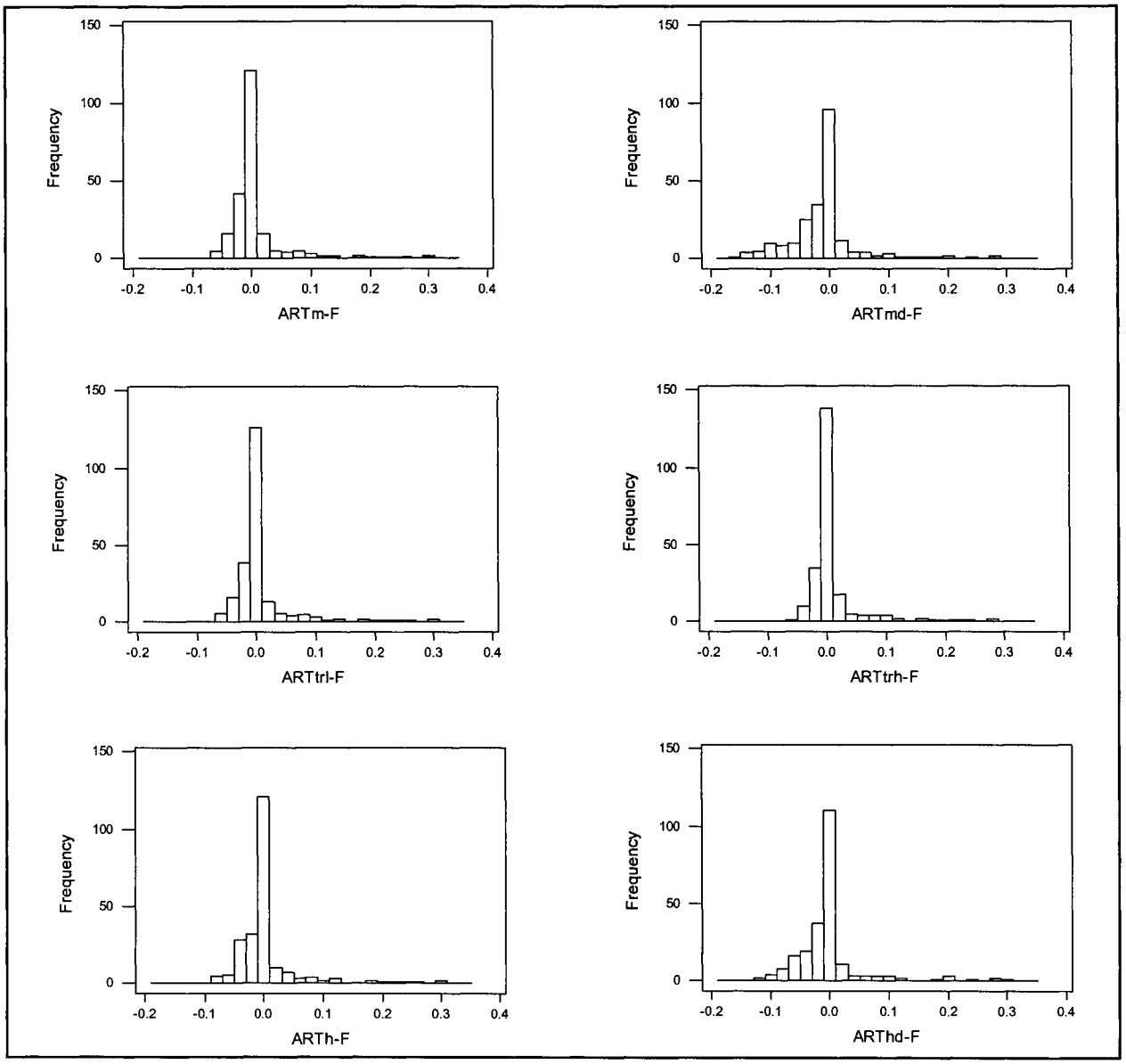

Figure 5. Histograms for Differences: (Alignment Statistic Power - F Statistic Power) for Condition 4, $\alpha = .05$.

Condition 4

Condition 4 modeled an ordinal interaction with two main effects. Although it took longer (more shift) to reach full power under this condition, many of the patterns observed with Condition 3 were observed here, too.

Figure 5, with histograms showing the power differences between the six aligning statistics and the F statistic, shows that, similar to Condition 3 results, the vast majority of differences are very close to zero. The data is skewed to the right, indicating cases where the alignment statistic is considerably more powerful than the F statistic, and the left tails (indicating a lack of power relative to the F) are the longest and heaviest for the ART_{md}, ART_{H} and ART_{HD}.

Table 2, which summarizes the differences between each of the alignment statistics and the F statistic used to construct the histograms in Figure 5, shows that the most extreme case of lack of power relative to the F test is with the ART_{HD} (-.1510), and the most extreme case of superior power relative to the F statistic is with the ART_{H} (.29880), with the ART_{m} next with .29700. A comparison of the differences by mean, median, minimum, maximum, Q1 and Q2 shows that the ART_{m}, ART_{trl}, and ART_{trh}, have a very slight advantage over the other alignment statistics.

Figures 6-8 show several situations where the lack of power of the F and ART_{md} are apparent. The F statistic showed a large deficiency in power for the Extreme Asymmetry data. The ART_{md} showed a deficiency for most of the distributions.

Conclusion

If Type I error is the major concern, the F statistic, ART_{md} and ART_{m}, in that order, were the most promising in this study. These statistics had no violations of a moderate definition of robustness, $\alpha \pm .25\alpha$, adjusted for sample size, for condition 1 and no violations of a stringent criterion, $\alpha \pm .1\alpha$, for Condition 2. The ART_{H} and ART_{HD} were the least satisfactory, with rates as high as 3.5α with the Extreme Asymmetry data when no main effects were present,

Table 2. Descriptive Statistics for Power Differences: (Alignment Statistic - F Statistic) for Condition 4, $\alpha = .05$.

Variable	N	Mean	Median	StDev	Minimum	Maximum	Q1	Q3
ART_{m}-F	229	0.00925	-0.0004	0.05363	-0.0638	0.2970	-0.0119	0.0045
ART_{md}-F	229	-0.01081	-0.0050	0.06102	-0.1510	0.2804	-0.0332	0.0000
ART_{trl}-F	229	0.00855	-0.0004	0.05293	-0.0648	0.2952	-0.0119	0.0035
ART_{trh}-F	229	0.01001	-0.0002	0.04803	-0.0530	0.2798	-0.0072	0.0053
ART_{H}-F	229	0.00425	-0.0012	0.05455	-0.0816	0.2988	-0.0160	0.0027
ART_{HD}-F	229	-0.00266	-0.0022	0.05773	-0.1192	0.2956	-0.0244	0.0004

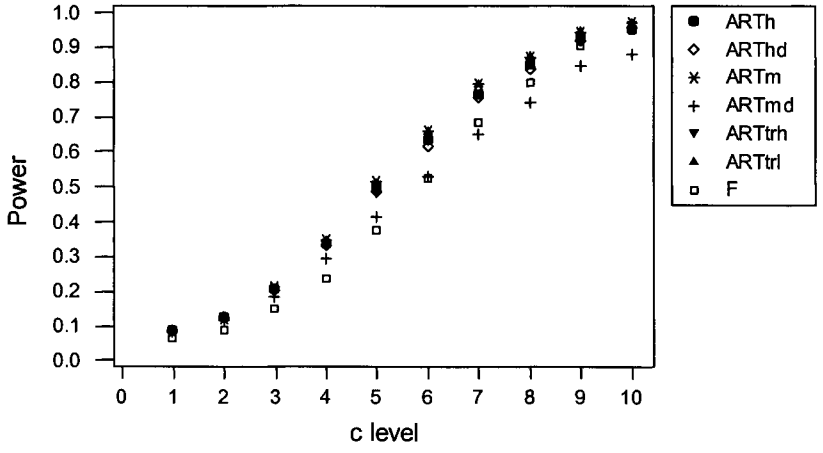

Figure 6. Power graph for Condition 4 (main effects with an ordinal interaction) for the Extreme Asymmetry data set with nominal alpha = .05 and n = 5. The c level is the multiple of $.25\sigma$ used for shift.

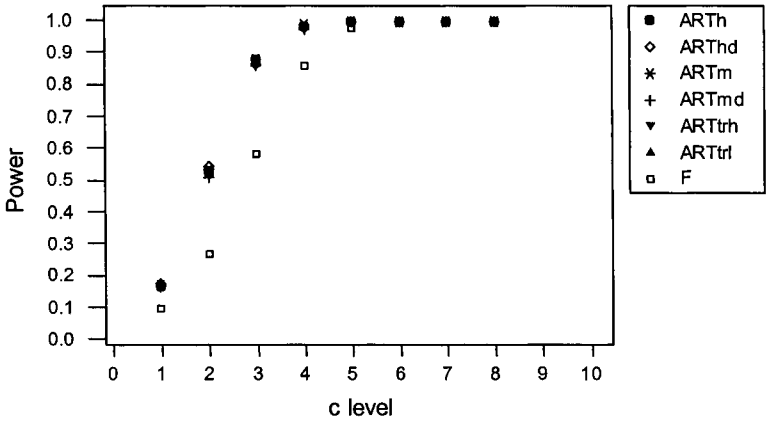

Figure 7. Power graph for Condition 4 (main effects with an ordinal interaction) for the Extreme Asymmetry data set with nominal alpha = .05 and n = 20. The c level is the multiple of .25σ used for shift.

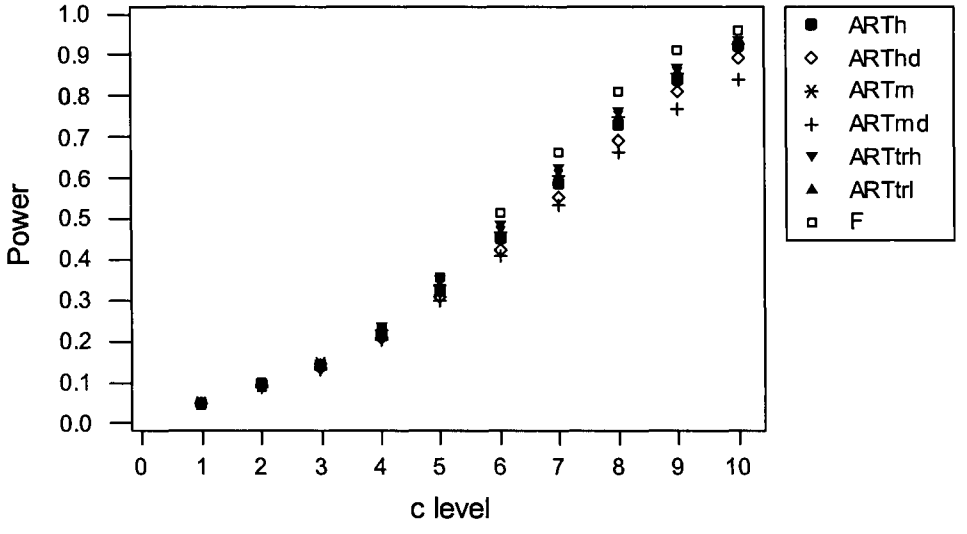

Figure 8. Power graph for Condition 4 (main effects with an ordinal interaction) for the Multi-modal and Lumpy data set with nominal alpha = .05 and n = 5. The c level is the multiple of .25σ used for shift.

and poor results with other distributions.

This study affirms the weak power of the F statistic, in comparison to the ART_m, as reported by Kelley and Sawilowsky (1997), with Extreme Asymmetry data, in a 2x2x2 design. The F statistic also performed poorly with the Likert data.

In addition, the median alignment showed lower power levels for most of the distributions studied, especially with Condition 4, an ordinal interaction with main effects. The good Type I error rates for the F statistic and the ART_{md} do not compensate for the much larger power deficiencies.

The ART_H and the ART_{HD} showed problems with both Type I error and power. The best statistics in terms of power for Conditions 3 and 4 were ART_{trh}, ART_{trl}, and the ART_m. The ART_m had a slight advantage in terms of Type I error rates; the two trims a slight power advantage.

Kelley and Sawilowsky (1997), in their study of the Blair-Sawilowsky test (which has been referred to in this study as ART_m) and other nonparametric tests for

interaction in a 2x2x2 layout came to this conclusion:

> It is recommended that when testing for interactions in a 2x2x2 layout, Analysis of Variance [F statistic] be used with data known to be symmetric with light tails, such as the normal and uniform distributions, and the Blair-Sawilowsky [ART_m] test be used with heavy-tailed or skewed data. If the shape of the distribution is unknown, the Blair and Sawilowsky test is recommended because it frequently exhibited considerably more power than the ANOVA [F]. In the apparently rare circumstances where data are obtained from a normal curve, this test will only be slightly less powerful than the ANOVA F test. (p. 357)

This study supports the value of the Blair-Sawilowsky (ART_m) and extending its application to a 3 x 4 layout. It also raises the possibility of other alignments (ART_{trl} and ART_{trh}) being as useful or even more so in other situations as mentioned above.

The F statistic has been considered an all-purpose statistic, used without consideration of the population. As has been indicated, this can lead to major errors. Although there is a natural tendency to want to find a substitute all-purpose statistic, there are many issues that would have to be addressed before any of these three could assume that role. Among them are: the nature of the interaction and number of non-null effects, other designs, the issue of unequal variances, and additional distribution issues.

Tukey (1984) described the practical power of a test as being the statistical power of a test multiplied by the probability that someone would actually use the test. This study has indicated three statistics as being somewhat equal for power and Type I error rates. Unless future studies indicate a big difference in the usefulness of the ART_{trl} and ART_{trh}, Tukey's criterion would favor the ART_m because it can be done quite easily on most statistical software packages. However, a macro for the Winsorized trimmed mean is available (Wilcox, 1996).

References

Blair, R. C., & Sawilowsky, S. S. (1990, April). *A test for interaction based on the rank transform.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Blair, R. C., Sawilowsky, S. S., & Higgins, J. J. (1987). Limitations of the rank transform statistic in tests for interactions. *Communications in Statistics-Simulation and Computation, 16*(4), 1133-1145.

Bradley, J. V. (1979). A nonparametric test for interactions of any order. *Journal of Quality Technology, 11*(4), 321-327.

Conover, W. J. & Iman, R. I. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician, 35*(3), 124-129.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*(30), 237-288.

Harrell, R. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika, 69,* 635-640.

Hettmansperger, T. F. (1984). *Statistical inference based on ranks.* New York: John Wiley.

Hoaglin, D. C., Mosteller, F. & Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis.* New York: John Wiley.

Kelley, D. L., & Sawilowsky, S. S. (1997). Nonparametric alternatives to the F statistic in analysis of variance. *Journal of Computer Simulations, 58,* 343-359.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

Micceri, T. (1986, Nov.). *A futile search for that statistical chimera of normality.* Paper presented at Florida Educational Research Association Annual Conference, Tampa, Florida.

Minitab, Release 12.1 (1998). State College, PA: Minitab, Inc.

Nanna, M. J., & Sawilowsky, S. S.(1998) Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods, 3*(1), 55-67.

Puri, M. L., & Sen, P. K. (1985). *Nonparametric methods in general linear models.* New York: Wiley.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60*(1), 91-126.

Sawilowsky, S. S., & Blair, R. C. (1987, April). *An investigation of the Type I error and power properties of the rank transform procedure in factorial ANOVA.* Paper presented at the annual meeting of the American Educational Research Association. Washington, DC.

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989), An investigation of the Type I error and power properties of the rank transform procedure in factorial anova, *Journal of Educational Statistics, 14*(3), 255-267.

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples t test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology, 60,* 240-243.

Shoemaker, L. H. (1986). A nonparametric method for analysis of variance. *Communications in Statistics*, *15*(3), 609-632.

Stigler, S. M. (1977). Do robust estimators work with *real* data? *The Annals of Statistics*, *5*(6), 1055-1098.

Toothaker, L. E., & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational and Behavioral Statistics*, *19*(3), 237-273.

Tukey, J. (1984). Untitled lecture at Woodrow Wilson Summer Program for Teachers, Princeton University, Princeton, NJ.

Wilcox, R. R. (1995). ANOVA: The practical important of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology*, *48,* 99-114.

Wilcox, R. R. (1996). *Statistics for the social sciences.* San Diego, CA: Academic Press.