Journal of Modern Applied Statistical Methods

Volume 12 | Issue 1 Article 7

5-1-2013

A Monte Carlo Simulation of the Robust Rank-Order Test Under Various Population Symmetry Conditions

William T. Mickelson
University of Wisconsin - Whitewater

Recommended Citation
Mickelson, William T. (2013) "A Monte Carlo Simulation of the Robust Rank-Order Test Under Various Population Symmetry Conditions," *Journal of Modern Applied Statistical Methods*: Vol. 12 : Iss. 1 , Article 7.
DOI: 10.22237/jmasm/1367381160

Journal of Modern Applied Statistical Methods
May 2013, Vol. 12, No. 1, 21-33

Regular Articles

A Monte Carlo Simulation of the Robust Rank-Order Test Under Various Population Symmetry Conditions

William T. Mickelson
University of Wisconsin – Whitewater
Whitewater, WI

The Type I Error Rate of the Robust Rank Order test under various population symmetry conditions is explored through Monte Carlo simulation. Findings indicate the test has difficulty controlling Type I error under generalized Behrens-Fisher conditions for moderately sized samples.

Key words: Robustness, hypothesis testing, Monte Carlo.

Introduction

Statistical significance tests are widely used in empirically based quantitative research and have been applied in virtually every field of study to test research hypotheses. Although many applied researchers use statistical methods, the pitfalls and limitations of statistical hypothesis tests due to violations of underlying assumptions are often overlooked (Kesselman, et al., 1998, Snyder & Thompson, 1998). It is known that the most commonly used statistical tests, the ANOVA F and Student's T-test, have underlying assumptions of independence of observations – that data are obtained from normally distributed populations having equal variances. Furthermore, these commonly used tests suffer potentially severe performance degradation when underlying assumptions of normality and equality of variance are not met (Glass, Peckham & Sanders, 1972). The violation of normality and equality of variance assumptions are often referred to as the generalized Behrens-Fisher problem.

When researchers are interested in testing the equality of two means – or medians – it is generally recommended that a nonparametric inference procedure, such as the Wilcoxon-Mann-Whitney test, be used (Harwell & Serlin, 1989). The Wilcoxon-Mann-Whitney test, however, is inappropriate for cases with unequal variances (Harwell, et.al. 1992; Zimmerman & Zumbo, 1993a, 1993b): It behaves much like the traditional student's t-test where the Type I error rate is depressed when the larger sample size is associated with the larger variance and, when the smaller sample size is associated with the larger variance, there is an inflation of the Type I error rate. Siegel and Castellan (1988) recommend that, in the case of non-normality and unequal variances, an alternative nonparametric inference procedure, the Robust Rank Order (RRO) test be used.

The Robust Rank Order (RRO) test (Fligner & Policello, 1981) is a modified version of the Mann-Whitney-Wilcoxon test designed to maintain both the nominal Type-I error rate and statistical power under generalized Behrens-Fisher conditions (Behrens,1929; Fisher, 1939; Scheffé, 1970; Zumbo & Coulombe, 1997). According to Siegel and Castellan (1988), the RRO test statistic approximates a normal distribution quickly as sample size increases; however, there is an underlying assumption of population distribution symmetry that is essential for the RRO test to be truly robust. Zumbo and Coulombe (1997) indicate this

William T. Mickelson is an applied statistician and consultant. Current interests include measurement, statistical robustness, modern re-sampling and nonparametric methods and the teaching and learning of statistics. Email him at: mickelsw@uww.edu.

quality is lacking under conditions of heterogeneity of variance.

Zumbo and Coulombe (1997) examined the performance of the RRO test for non-normal populations with unequal variances but confined their study to small sample size cases (n between 3 and 12) where the exact level of the RRO test was known. They found that the RRO test was conservative for symmetric distributions and that it performed inconsistently when the population distribution was skewed. Vargha and Delaney (2000) also conducted a simulation study to examine the performance of the RRO test, the results of which conflicted with Zumbo and Coulombe.

In this study, the Type I Error Rate of the RRO test is examined for moderately sized samples. Sample sizes examined are larger than those for which an exact test is possible, yet potentially before the asymptotic convergence to the standard normal distribution has occurred. This work extends Zumbo and Coulombe's (1997), and Vargha and Delaney's (200) simulation research on the RRO test to a larger range of variance inequality and non-normality situations, and sample sizes that are typically found in researcher practice.

The Robust Rank Order Test

Let X1, X2, …, Xm and Y1, Y2, …, Yn denote two independent random samples from parent populations with continuous distribution functions F(X) and G(Y), respectively. If it is assumed that a treatment effect will manifest itself as a difference in the location of the experimental group's location, then the null hypothesis for the RRO test is:

$$\mathrm{H_0: Median(X) = Median(Y)}$$
$$\text{versus}$$
$$\mathrm{H_a: Median(X) \neq Median(Y)} \tag{1}$$

The RRO test is a distribution-free test of (H_a). The following steps are used to compute the RRO test:

1. For each observation in group 1, let Pi = [Number of observations in group 2 < Xi], for i = 1,2,…,m.

2. For each observation in group 2, let Qj = [Number of observations in group 1 < Yj], for j = 1, 2, …., n. The Pi and Qj are called the placements.

3. Compute the average of the placements for group 1 and group 2, termed $\bar{P}$ and $\bar{Q}$, respectively.

4. Compute the sum of squares of placement deviations for each group. The formulas for these computations are:

$$V_1 = \sum_{i=1}^{n}\left(P_i - \bar{P}\right)^2$$

and

$$V_2 = \sum_{j=1}^{m}\left(Q_i - \bar{Q}\right)^2$$

5. The Robust Rank Order test statistic is:

$$TS = \frac{\left(n*\bar{P}\right)-\left(m*\bar{Q}\right)}{2\sqrt{V_1+V_2+\bar{P}*\bar{Q}}}$$

Fligner & Policello (1981) give critical values for small sample sizes. For larger sample sizes, the test statistics TS is distributed as a standard normal distribution.

Methodology

Monte Carlo simulation was used to estimate Type I error rates for the Robust Rank Order (RRO) Test under various population symmetry conditions. All individual estimates are based on 20,000 iterations. Three replications per condition were obtained in order to model and graph the Type I error rates as response surfaces. The conditions modified in the simulation consisted of:

- Total Sample Size (N = 30, 50, 100, 150)

- Unequal sample sizes per group in terms of sample size ratios of:

  - 1:1
  - 1.5:1
  - 2.3:1
  - 4:1
  - 9:1

- Variance ratios ranging from 1:1 to 20:1

- Inverse and direct pairing of sample size with variances

- Population distributions consisting of:
  - Symmetric Normal
  - Symmetric Uniform
  - Symmetric T-distribution with df=3 (heavy tailed)
  - Moderately skewed (Weibull with parameters a=2 and b=2)
  - Heavily skewed (Weibull with parameters a=1.5 and b=1)

Type I error estimation was defined and calculated as the number of times the test statistics for both tests rejected the null hypothesis divided by the maximum number of iterations (20,000) when the null hypothesis is true. GAUSS programming language was used to run the simulation.

Consistent with recommendations of Bradley (1978) for evaluating Type I error rate estimates, multiple benchmarks for the criteria of robustness were used. Specifically, Type I error rate estimates between $\alpha \pm \alpha/10$, or [0.045, 0.055] for a nominal 0.05 level test, were considered robust at a stringent level. Other benchmarks used include: a) intermediate level, $\alpha \pm \alpha/4$; b) liberal, $\alpha \pm \alpha/2$; and c) very liberal, $\alpha \pm 3\alpha/4$. Graphical representations of the data are presented for selected conditions to illustrate the primary findings of this simulation study. A database of the Type I Error rate estimates under all of the simulated conditions is available from the author (mickelsw@uww.edu). Selected graphical representations of the data are presented herein.

Results

Type I error rate results are organized by the combinations of sample size paired with variance ratios.

Test Performance: Equal Sample Sizes per Group

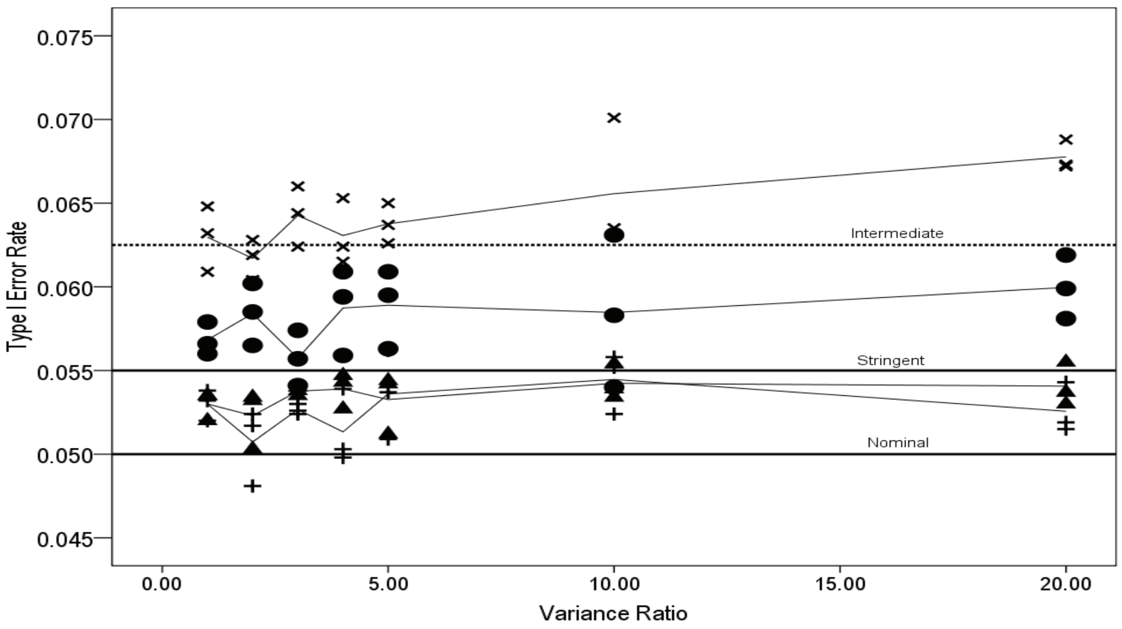
Figures 1 - 5 present the results of the Robust Rank Order (RRO) test when sample sizes are equal under 5 different population symmetry situations. The X-axis in these plots corresponds to increasing levels of population variance heterogeneity, where the variance ratio ranges from 1:1 to 20:1. The Y-axis corresponds to the Type I Error rate. The results delineated by total sample size. The lines in the graph relative to each total sample size level represent interpolation between means.

As shown in Figures 1, 2 and 3 the RRO Test is relatively robust for the symmetric population distribution conditions and sample sizes under consideration. All Type I error rates are within liberal robustness standards, and within intermediate robustness standards for total sample sizes of 50 and larger. There is some Type I Error rate inflation for the smaller sample size. Not surprisingly, the error rates improve with increasing total sample size.

Figure 4 illustrates that the RRO is also relatively robust under the condition of the moderately skewed population distribution. Total sample sizes of 50 or larger result in error rate estimates within the intermediate robustness standard, however the smallest sample size under consideration does experience Type I Error Rate inflation approaching the very liberal standard, furthermore the error rate appears to increase as the variance ratio increases. Figure 5 demonstrates that the RRO is not robust to moderately extreme asymmetry in the population distribution. Further analysis indicated that this phenomenon was prevalent under all conditions evaluated when the population distribution was heavily skewed, as such, this distribution is not presented.

In sum, the RRO test can be considered robust when sample sizes are equal and the population distribution can be considered symmetric or at worst, moderately skewed. For small total sample size (N=30), the RRO exhibits some Type I Error rate inflation, however sample sizes of 50 or larger have Type I Error rate controlled at the intermediate level of robustness.

Figure 1: RRO Test Type I Error Rate – Normal Distribution



Sample Size

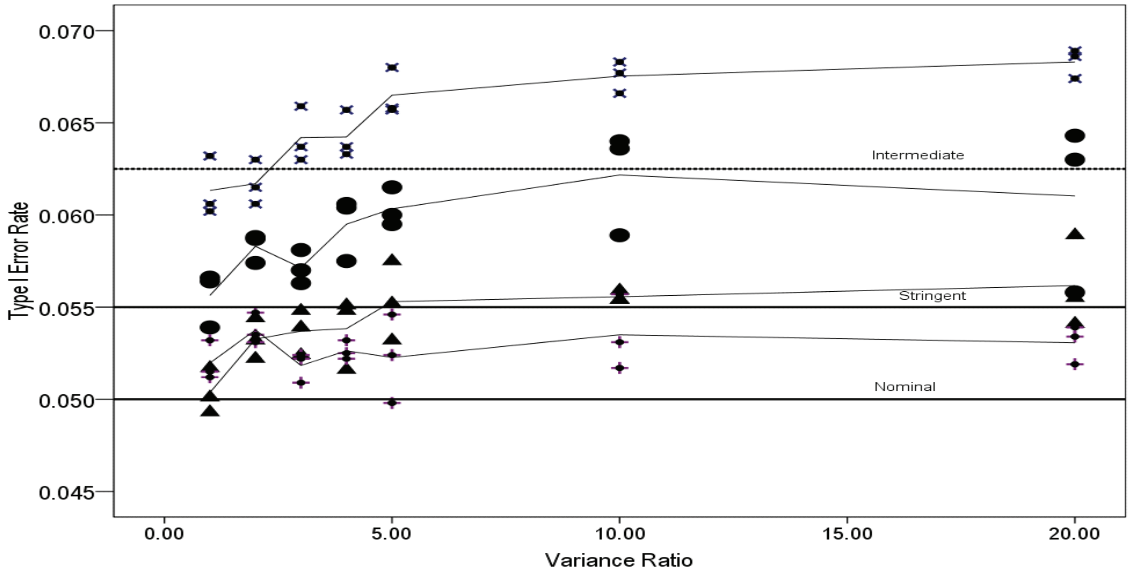
X N=30

? N=50

? N=100

+ N=150

Figure 2: RRO Test Type I Error Rate – Uniform Distribution



Sample Size

X N=30

? N=50

? N=100

+ N=150

Figure 3: RRO Test Type I Error Rate – Heavy-Tailed Distribution

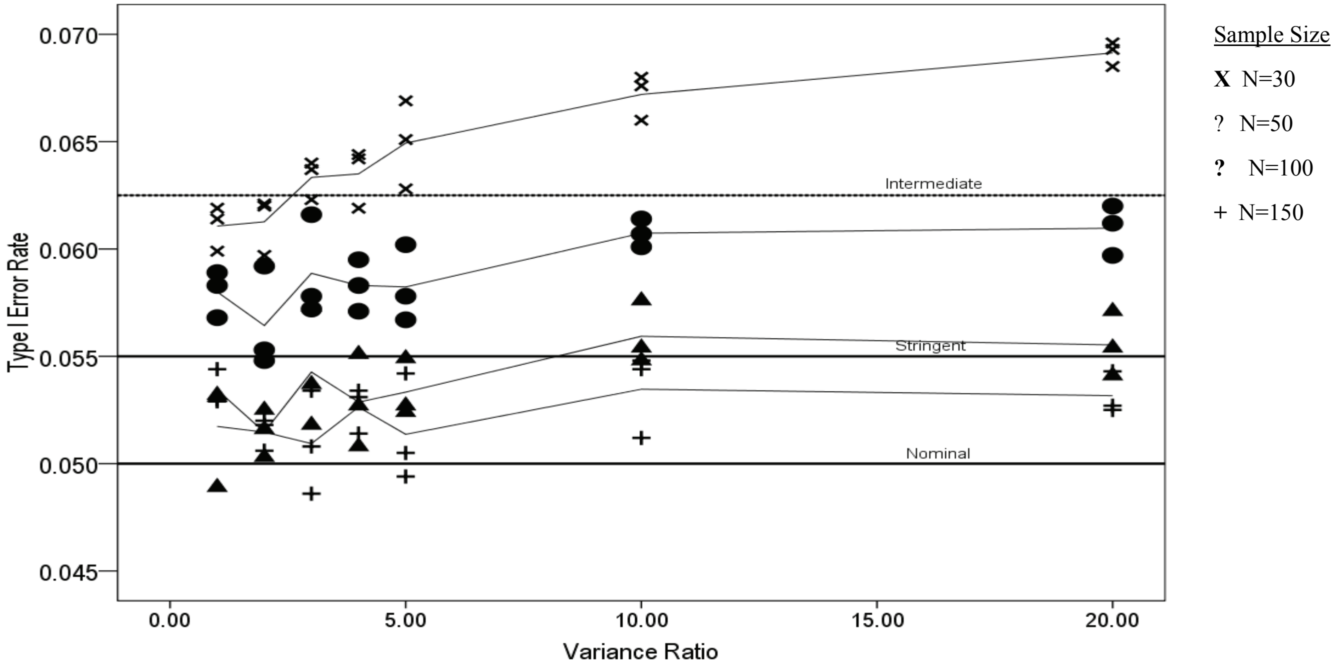


Figure 4: RRO Test Type I Error Rate – Moderately Skewed Distribution

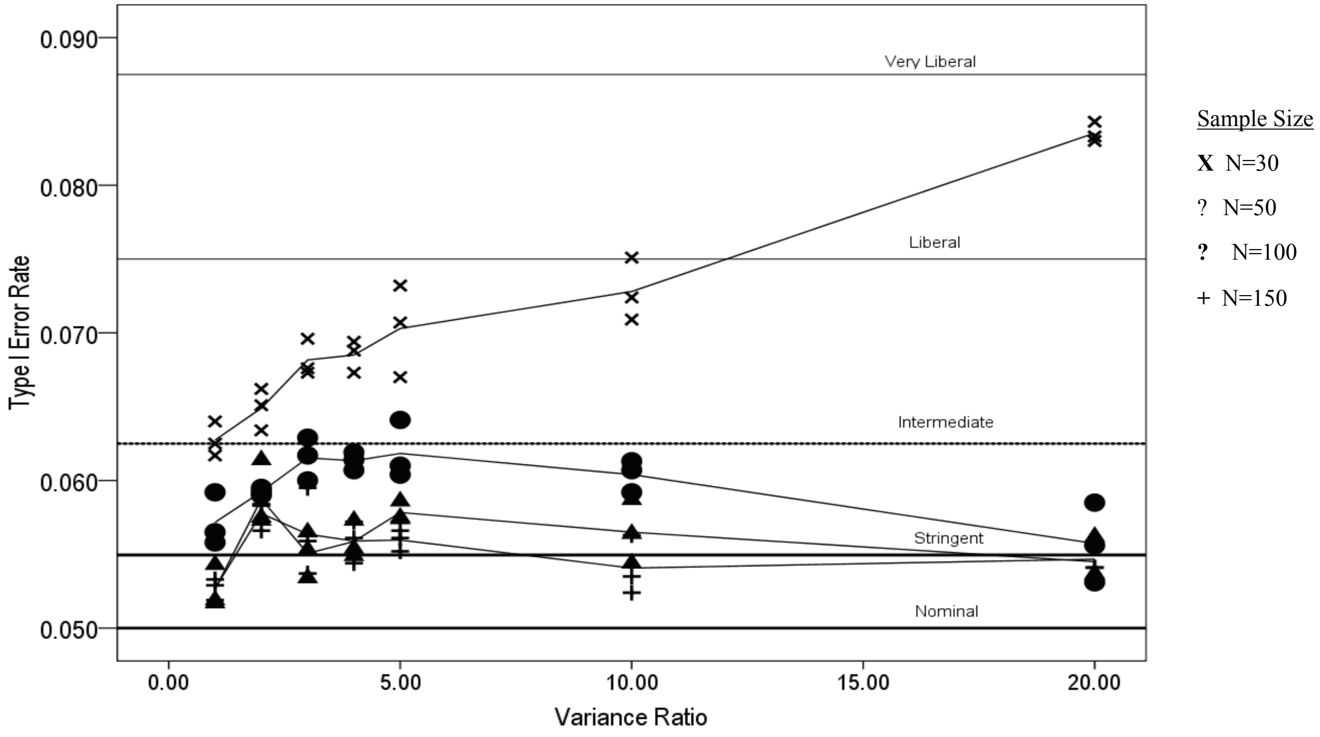
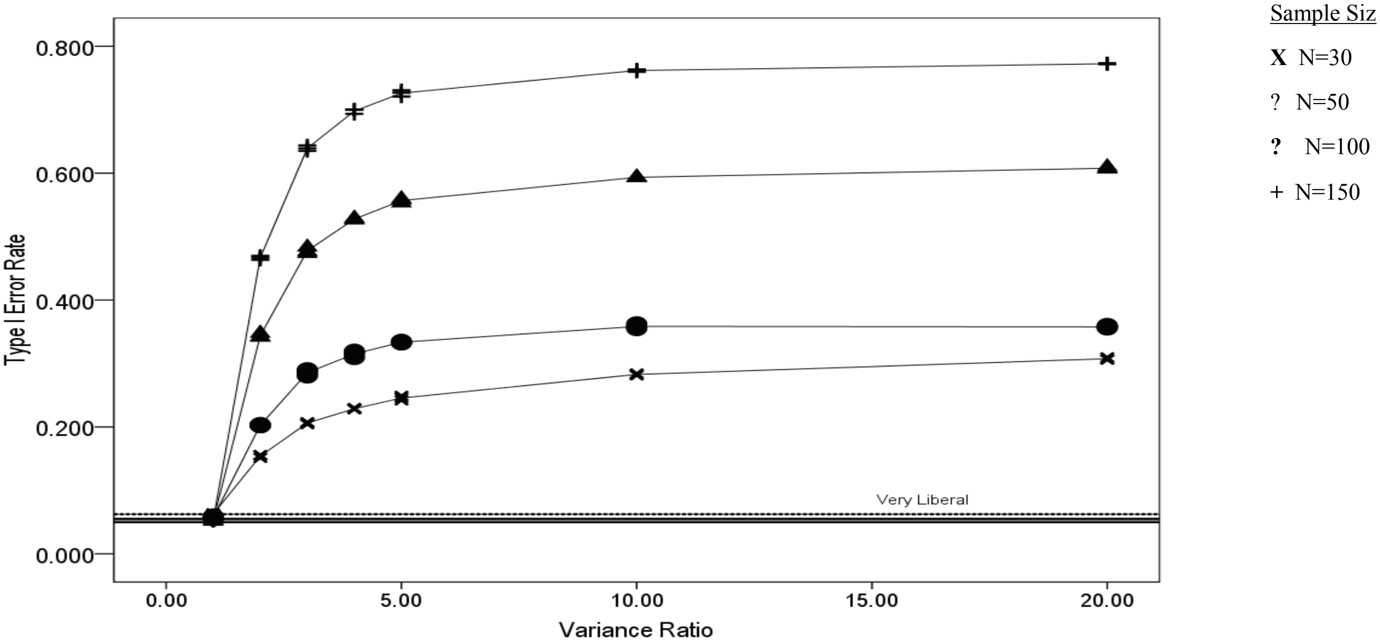


Figure 5: RRO Test Type I Error Rate – Heavily Skewed Distribution



Test Performance: Inverse Pairing

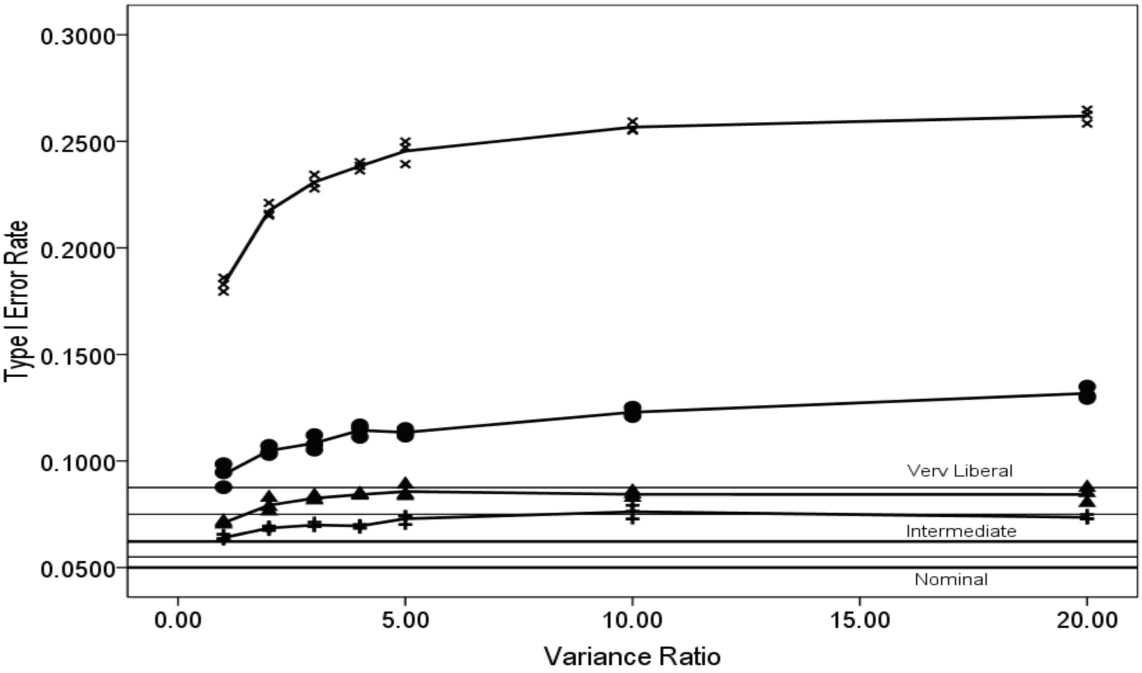
Selected Type I Error results for the Robust Rank Order (RRO) test under the situation of inverse pairing (sample sizes in each group are unequal and the group with the largest sample is paired with the smallest population variance) are presented in Figures 6 - 11. The X-axis in these plots corresponds to increasing levels of population variance heterogeneity, where the variance ratio ranges from 1:1 to 20:1. The Y-axis corresponds to the Type I Error rate. The degree of sample size inequality is given in the legend of the graph; lines in the graph represent interpolation between means within levels of the sample size ratio.

Figures 6 and 7 present the Type I error rate results for the RRO test under the situation of inverse pairing when the population is normally distributed for total overall sample size of 30 and 150, respectively. Figure 6 shows that when the total sample size is small, N=30, the RRO is not robust under inverse pairing when the sample size ratio is 4:1 or larger. In general, Type I Error Rate inflates as the sample size ratio and variance ratios increase, either independently or together. Across Figures 6 and 7, the situation generally improves as the total sample size increases. Figure 7 shows that when N=150, RRO test appears to be robust at the moderate robustness standard or better provided the sample size ratio is 4:1 or less. The RRO cannot be considered robust when there is great discrepancy between the sample sizes and the sample size ratio is 9:1 or larger. When the overall sample size increases to N=150, the RRO test is robust at the most liberal robustness standard for the largest sample size ratio and can be considered robust at the intermediate robustness standard for sample size ratios of 4:1 or smaller.

In examining the Type I Error Rate performance when the population has a uniform distribution, the pattern to the results is highly similar to the patterns observed when the population is normally distributed. As such, those findings are not summarized here and the reader is referred to the Type I Error Rate database available by request from the author.

Figures 8 - 11, present the Type I error rate results for the RRO test under the situation of inverse pairing when the population distribution is heavy tailed, and moderately skewed, respectively. Results are presented for

Figure 6: RRO Test Type I Error Rate – Normal Distribution N=30



Sample Size Ratio

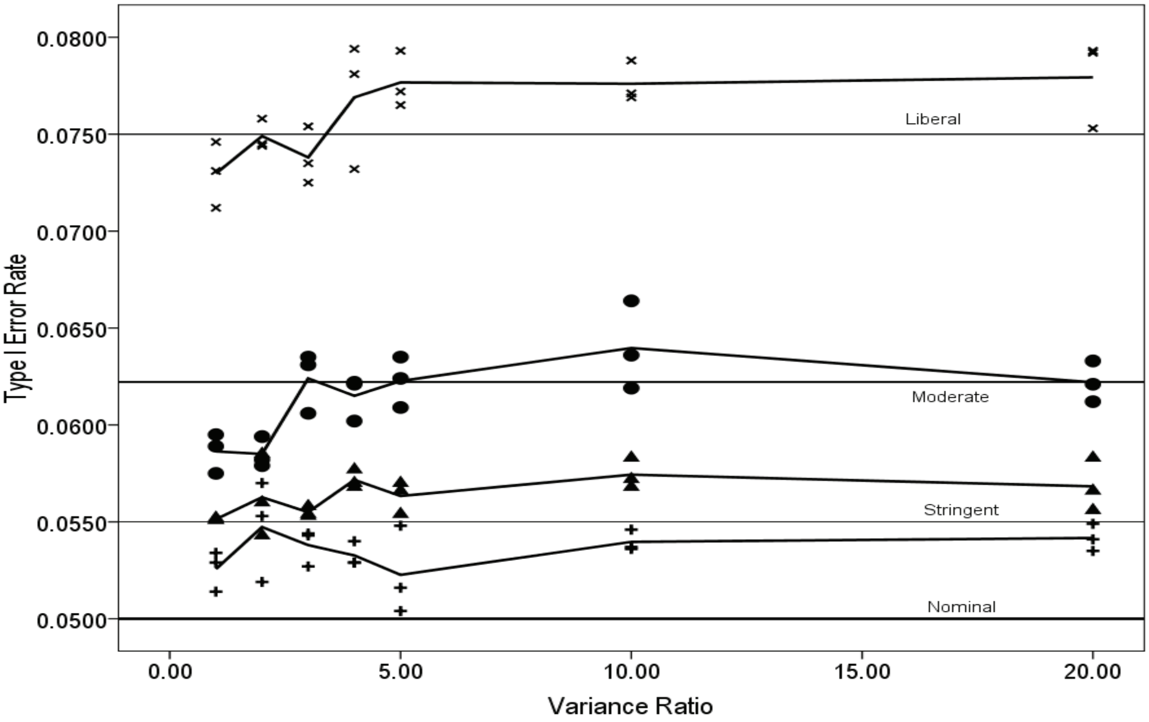
X 9.0 : 1

? 4.0 : 1

? 2.3 : 1

+ 1.5 : 1

Figure 7: RRO Test Type I Error Rate – Normal Distribution N=150



Sample Size Ratio

X 9.0 : 1

? 4.0 : 1

? 2.3 : 1

+ 1.5 : 1

total overall sample sizes of 30 and 150, respectively. Type I Error Rate estimates for intermediate sample sizes and additional population distributions are available by request from the author.

The Type I Error Rate estimates for the heavy tailed (Figures 8 and 9) and moderately skewed, (Figures 10 and 11) population distributions demonstrate patterns that are very similar to those previously observed when the population distribution was normally distributed. The RRO test does not perform well under inverse pairing for small sample sizes and performance degrades as the discrepancy in sample size increases, that is, increasing sample size ratio. Performance does improve with increasing sample size, however, even with sample sizes as large as N=150 the higher the discrepancy in sample size situations still have not become sufficiently robust to claim Type I Error rate is controlled. The RRO performs less well when the population distribution is moderately skewed. The same general patterns as those of the symmetric distributions emerge, however, with Type I Error rates becoming more controlled and robust as the sample size ratio becomes closer to 1:1.

The claim by Fligner and Policello (1981) that the RRO test statistic converges to a standard normal distribution with increasing sample size for symmetric population distributions appears to be confirmed from this evidence. Even at the sample size of N=150, however, the convergence has not fully materialized to warrant calling the RRO test truly robust at the most stringent level of robustness for heavily-tailed symmetric distributions. There is some evidence that the RRO test could be used in situations with data from moderately skewed population distributions, but only in cases where the sample size ratios are 4:1 or less.

Test Performance: Direct Pairing

Type I Error results for the Robust Rank Order (RRO) test under the situation of direct pairing (unequal sample sizes between groups and the group with the largest sample is paired with the largest population variance) when the populations are normally distributed. Direct pairing occurs when population variances are unequal and the group with the smallest variance is paired with the smallest sample size. The X-axis in these graphs corresponds to increasing levels of population variance heterogeneity, where the variance ratio ranges from 1:1 to 20:1. The Y-axis corresponds to the Type I Error rate. The degree of sample size inequality is measured by the ratio of the sample sizes and is given in the legend of the graph; lines in the graph represent interpolation between means within levels of the sample size ratio.

Figures 12 and 13 present the Type I error rate results for the RRO test under the situation of direct pairing when the population is normally distributed for total overall sample sizes of 30 and 150, respectively. Figure 12 shows that when the total sample size is small, N=30, the RRO test is predominantly robust and controls the Type I Error Rate rather well. There is one exception, however, in that the RRO test is not robust under direct pairing when the sample size ratio is 9:1 and there is some indicate that there are some issues with error rate control when the sample size ratio is 4:1. In general, as sample size ratio increases, the RRO test becomes less robust and this pattern is more pronounced as the sample size ratio exceeds 4:1. As the overall sample size increases, however, the RRO is robust for the direct pairing situation (see Figure 13).

Interestingly, the Type I Error Rate patterns exhibited by the RRO when the population is normally distributed are the same for the heavy-tail and moderately skewed distributions. Lack of robustness occurs with the low overall sample size, N=30, and dramatically improves with increasing N. When overall sample sizes is large, N=150, the RRO test can be considered robust, even for the moderately skewed distribution (these findings are not illustrated here and the reader is referred to the Type I Error Rate database available by request from the author. Only for the heavily skewed population distribution does the RRO fail to control Type I error rate well for the direct pairing situation when sample size is large.

Figure 8: RRO Test Type I Error Rate – Heavy Tail Distribution N=30

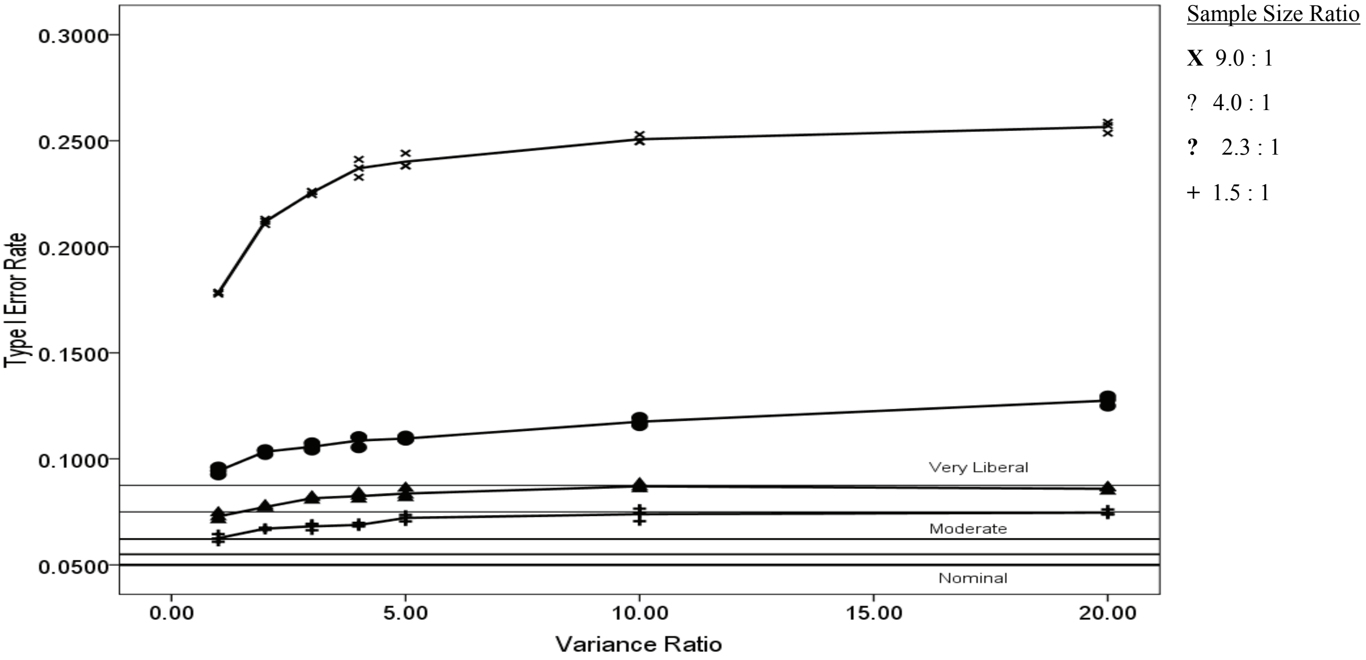


Figure 9: RRO Test Type I Error Rate – Heavy Tail Distribution N=150

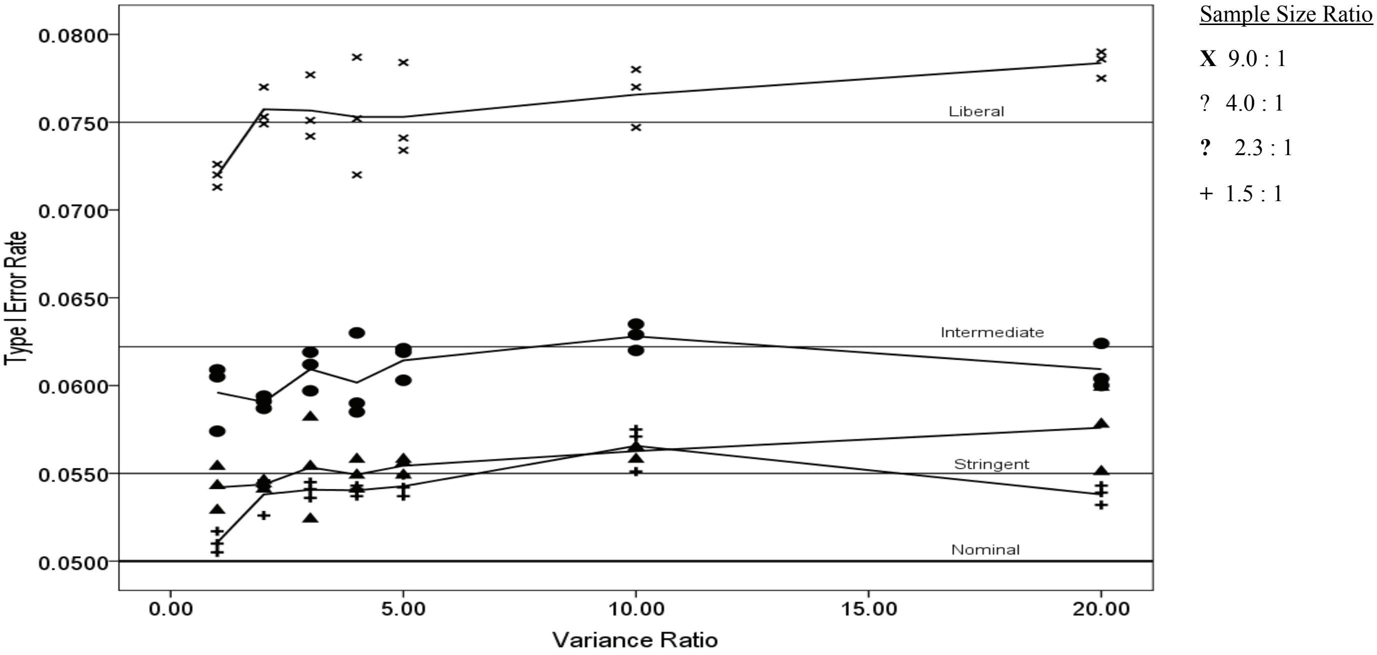


Figure 10: RRO Test Type I Error Rate – Moderately Skewed Distribution N=30

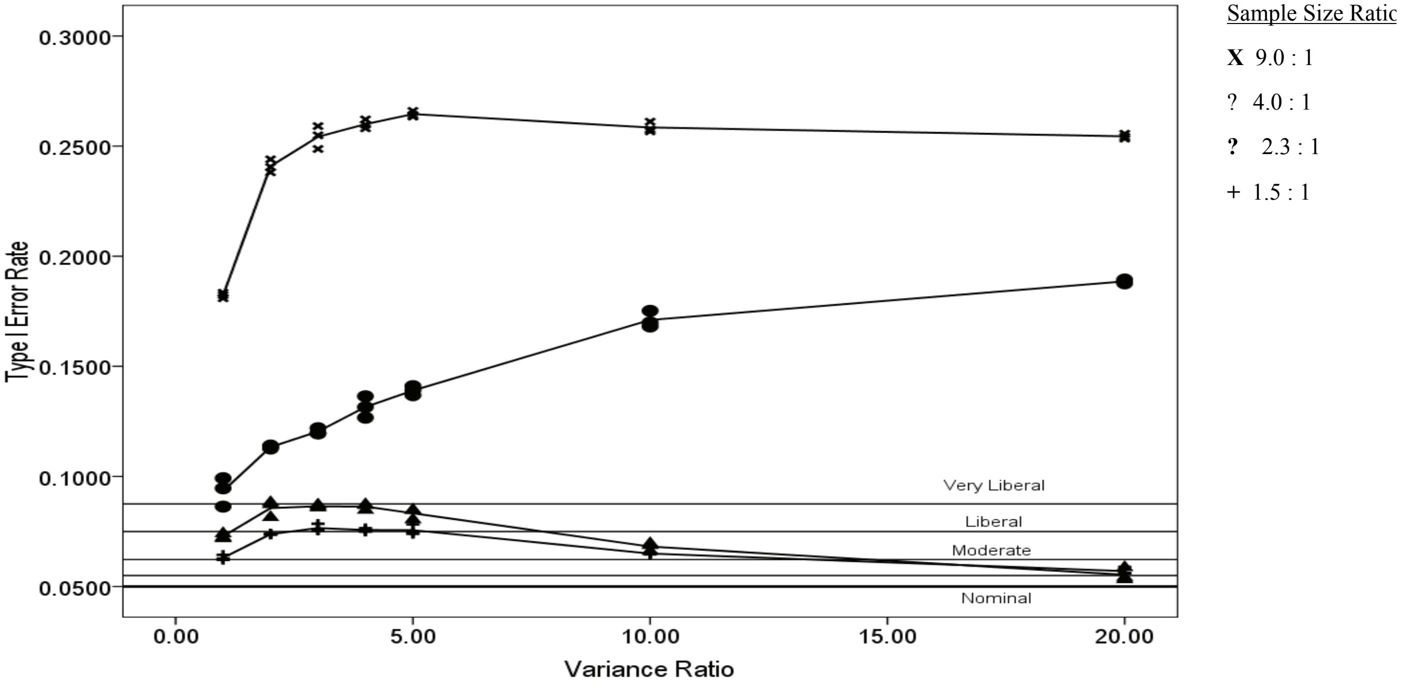


Figure 11: RRO Test Type I Error Rate – Moderately Skewed Distribution N=150

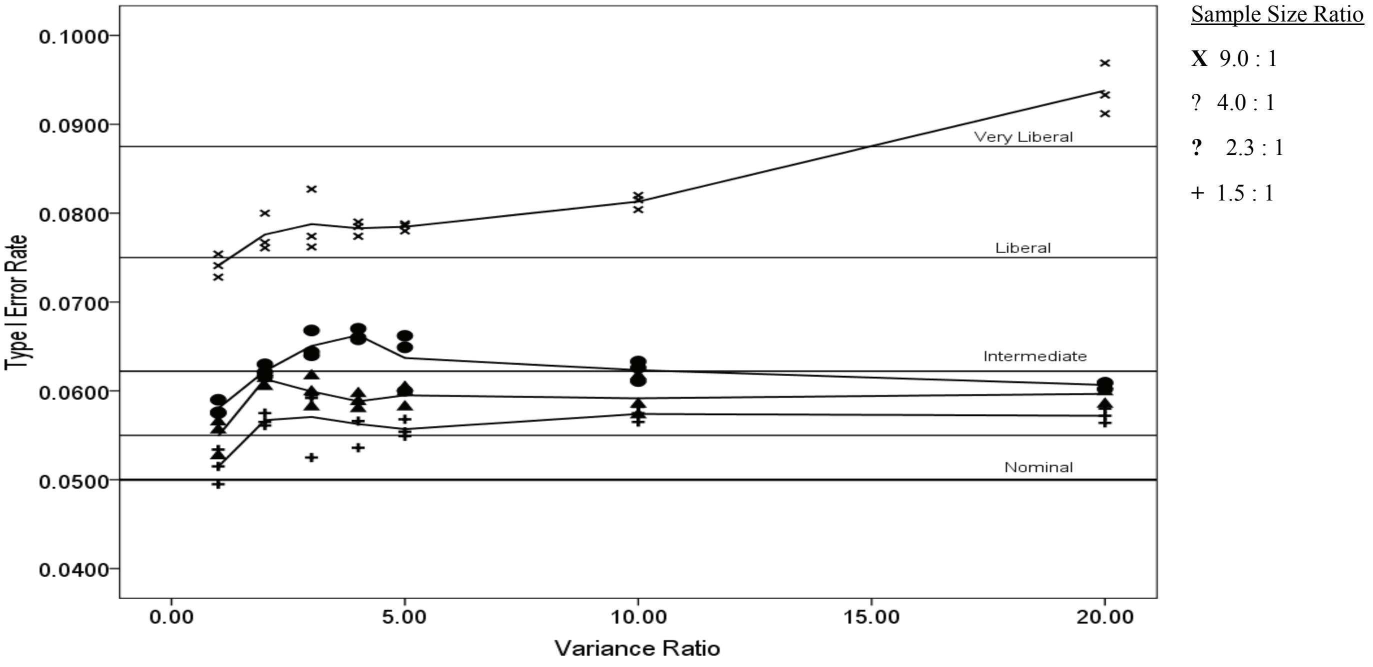


Figure 12: RRO Test Type I Error Rate – Normal Distribution N=30

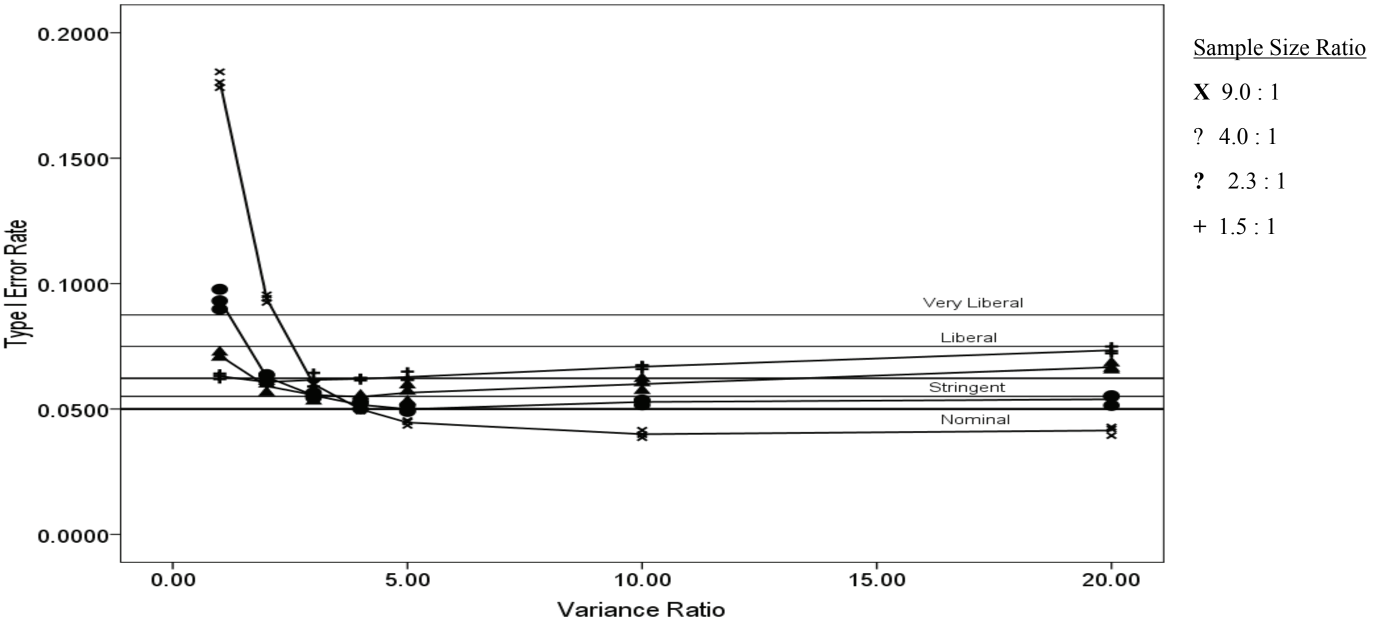
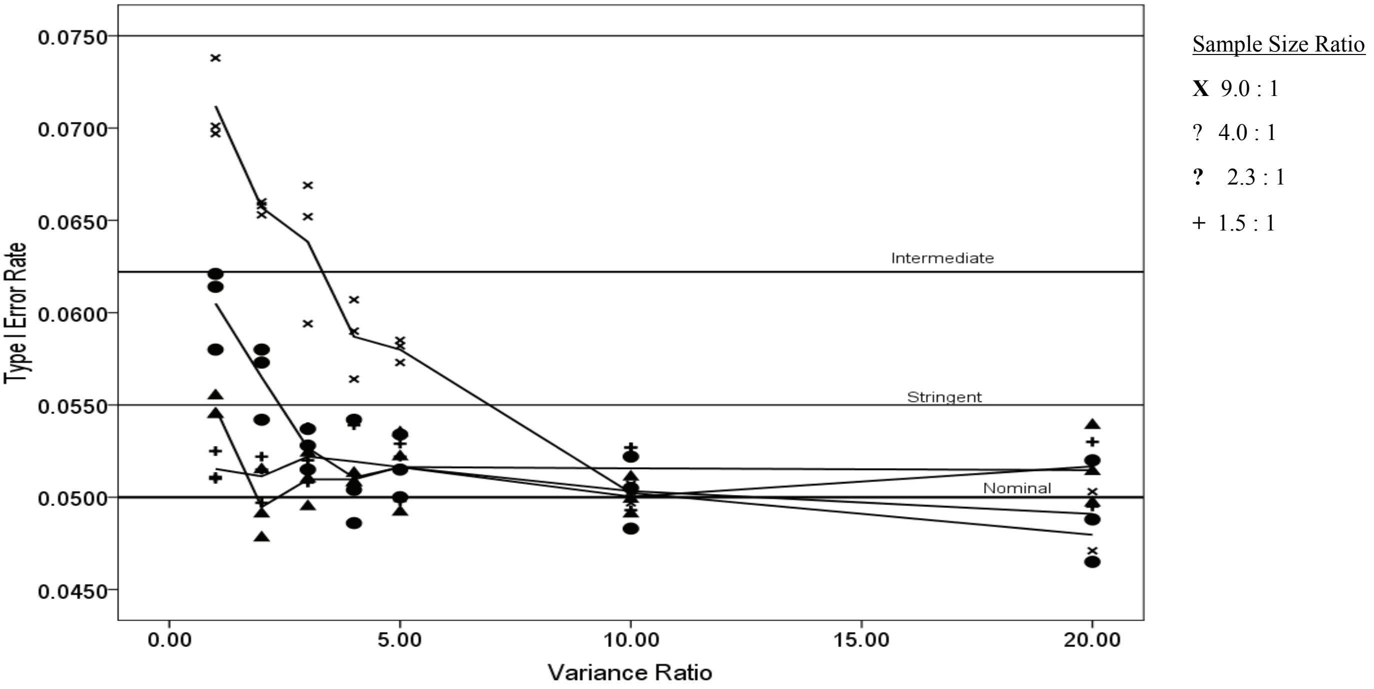


Figure 13: RRO Test Type I Error Rate – Normal Distribution N=150



Conclusion

This study investigated the performance of the Robust Rank Order (RRO) test (Fligner & Policello, 1981) under various population symmetry conditions in the intermediate sample size range prior to the asymptotic distribution holding. First, the claim by Fligner and Policello (1981) that it is necessary to assume the underlying population distributions are symmetric is confirmed. However, to a modest degree, the RRO test does control the Type I Error for moderately skewed population distributions. In general, it appears that the RRO test has the tendency to be liberal, with Type I error rate estimates becoming increasingly inflated:

- as the population distribution becomes more skewed;
- when variance ratios and/or sample size ratios become larger; and
- when overall sample size is smaller.

Particularly interesting is the finding that the RRO appears to perform better under the situation of direct pairing than it does when inverse pairing is present.

When sample sizes are equal, the RRO test controls Type I error rate at essentially the nominal level. The RRO has a slightly inflated Type I error rate, but for the symmetric population distributions, this rate inflation is moderate, at worst, and performance improves with increasing sample size. Under inverse pairing, the RRO test does not perform particularly well in controlling Type I error rate. There is considerable rate inflation that increases as the sample size ratio and/or variance ratio's increase. Performance does improve with increased sample sizes, but even when total sample size reaches N=150 the Type I Error rate is not fully controlled; this is particularly true the more sample sizes become disparate. For smaller sample sizes the RRO test cannot be recommended under conditions of inverse pairing. However, for the direct pairing situation the RRO performs moderately well in controlling Type I error rate. There is some rate inflation for the large sample size and variance ratios when overall sample size is small, but performance does improve dramatically with increased sample size.

In summary, the RRO test, billed as a statistical test designed to maintain the nominal Type-I error rate under generalized Behrens-Fisher conditions, did not perform uniformly well. The RRO improves with increasing sample size, but has a difficult time with inverse pairing of sample size and variance inequality. Overall, results of this study indicate that the asymptotic result of the RRO test has not sufficiently come into play when overall sample size is between N=30 and N=150 to make the test uniformly robust. The RRO test can be cautiously used in these overall sample size ranges, provided the sample size ratios are less than 4:1 and it can be reasonably assumed that the population distribution is symmetric or – at worst – moderately skewed.

References

Behrens, W. V. (1929). Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. *Landwirtsch, Jahrbucher*, *68*, 807-837.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Fisher, R. A. (1939). The comparison of samples with possibly unequal variances. *Annals of Eugenics*, *9*, 174-180.

Fligner, M. A., & Policello, G. E. II (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association*, *76*, 162-168.

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*(*3*), 237-288.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*, 315-339.

Harwell, M. R., & Serlin, R. C. (1989) A nonparametric test statistic for the general linear model. *Journal of Educational Statistics*, *14*, 351-371.

Keselman, H.J., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*(*3*), 350-386.

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association*, *65,* 1501-1508.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences.* New York, NY: McGraw-Hill.

Snyder, P. A., & Thompson, B., (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly*, *13*(*4*), 335-348.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*, 101-132.

Zimmerman, D. W., & Zumbo, B. D. (1993a). Rank transformations and the power of the Student t test and Welch t test non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology*, *47*, 523-539.

Zimmerman, D. W., & Zumbo, B. D. (1993b). The relative power of parametric and nonparametric statistical methods. In *A handbook for data analysis in the behavioral sciences: Methodological issues*, G. Keren & C. Lewis (Eds.), 481-517. Hillsdale, NJ: Erlbaum.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology*, *51*, 139-149.