Journal of Modern Applied Statistical Methods

Volume 8 | Issue 1 Article 3

5-1-2009

Comparative Power Of The Independent t, Permutation t, and WilcoxonTests

Michèle Weber
Private Scholar, mi.fatal-weber@att.net

Shlomo Sawilowsky
Wayne State University, shlomo@wayne.edu

Journal of Modern Applied Statistical Methods
May 2009, Vol. 8, No. 1, 10-15

1538 – 9472/09/$95.00

Comparative Power Of The Independent t, Permutation t, and WilcoxonTests

Michèle Weber
Private Scholar

Shlomo Sawilowsky
Wayne State University

The nonparametric Wilcoxon Rank Sum (also known as the Mann-Whitney U) and the permutation t-tests are robust with respect to Type I error for departures from population normality, and both are powerful alternatives to the independent samples Student's t-test for detecting shift in location. The question remains regarding their comparative statistical power for small samples, particularly for non-normal distributions. Monte Carlo simulations indicated the rank-based Wilcoxon test was found to be more powerful than both the t and the permutation t-tests.

Key words: t test, Wilcoxon, permutation, power.

Introduction

When testing for shift in location, Blair and Higgins (1985b) and Sawilowsky (1992; see also 1990) demonstrated that the nonparametric Wilcoxon Rank Sum test (also known as the Mann-Whitney U) is more powerful than the two independent samples Student's t test for data obtained from non-normal populations. For example, the Wilcoxon test can be up to four times more powerful than the t-test when the data are sampled from an exponential distribution (Sawilowsky & Blair, 1992).

Permutation techniques are also distribution-free (Bradley, 1968; Edgington, 1995; Maritz, 1981; Mielke & Berry, 2001). In this context, they require independence (Good, 1994; Maritz, 1981), exchangeability (Boik, 1987; Commenges, 2003; Good, 2002), continuity of the distributions (Edgington, 1995), and homogeneity of variance (Boik, 1987). Regarding their power properties, Good (1994), among many other authors, postulated that permutation methods are superior in terms of comparative power as compared with nonparametric procedures.

Michèle Weber is a private scholar in San Jose, California. Email: mi.fatal-weber@att.net. Shlomo Sawilowsky is a professor of educational statistics, and editor of JMASM. Email: shlomo@wayne.edu.

Adams and Anthony (1996) and Ludbrook and Dudley (1998) agreed with this view, and asserted that the reason permutation tests have higher power than nonparametric counterparts is because of the use of actual data instead of ranks. However, in a Letter to the Editor published in *The American Statistician*, Higgins and Blair (2000) demurred, and countered that statistical power is not lost via ranking data.

The same point was made previously by Blair (1985), "I have never seen an assertion of parametric power superiority accompanied by a citation to support the position. This is not too surprising since the statistical literature does <u>not</u> support such a position" (p. 4-5). This sentiment was echoed by Sawilowsky (1993) via an analogy:

> Both an accomplished opera singer sings and an off-key beginning tuba player plays dots and dashes of the International Morse code. While some may consider the opera singer's notes to be sounds of music, there is, in fact, no more information in those dots and dashes than in the off-key notes of the beginning tuba player, with respect to the code. If the complexity and subtlety of what is often imagined to be included in interval scales is noise and not

> signal, parametric tests will have no more information available than a rank test, and will be less efficient by trying to discriminate a signal from noise when in fact there isn't any. (p. 398)

Purpose of the study

Higgins and Blair (2000) opined that the Wilcoxon test is more powerful than the permutation t-test (and Student's t-test) when testing for shift in location. They postulated that the power properties of the permutation statistic follow the spectrum of the native test, not the nonparametric alternative. The purpose of this study, therefore, is to determine if indeed the permutation t-test follows the power properties of the two independent samples Student's t, or if it is fact superior to the nonparametric Wilcoxon Rank Sum test.

The resolution of this debate will have considerable impact on real data analysis with small samples in applied research. The rationale for selecting an optimum method for statistical analysis resides in the importance of detecting a treatment effect or naturally occurring condition, even it is subtle, assuming that it exists. The ability to detect the effect is quantified by the statistical power of the test. This makes the study of the comparative power properties of the permutation technique very important in applied research, where the effect size of treatments or interventions is oftentimes very small.

Methodology

A Fortran program was written to study the properties of the two independent samples Student's t test, the permutation t test, and the Wilcoxon Rank Sum test. Nominal alpha was set to $\alpha = 0.05$. The sample sizes studied were $n_1 = n_2 = 10$; $n_1 = 5$, $n_2 = 15$; $n_1 = n_2 = 20$; and $n_1 = 10$, $n_2 = 30$. Data were drawn from a normal distribution ($\mu = 0$, $\sigma = 1$), exponential distribution ($\mu = \sigma = 1$) and Chi-square distribution (df = 1).

The Type I error portion of the study was conducted by drawing samples with replacement for the various combinations of sample sizes and distribution, conducting the hypothesis tests, recording the results, and repeating the experiment for one million repetitions per study parameter. The power portion of the study was based on 1,500 repetitions per experiment. The reduction in repetitions was required due to the CPU time necessary for permutation intensive computations. The means were shifted by $\mu = .2\sigma$, $.5\sigma$, $.8\sigma$, and 1.2σ of the respective distribution.

Results

Type I Error Rates

The Type I error rates, which have been extensively studied elsewhere, are briefly repeated here to demonstrate the veracity of the Fortran program. All Type I error results replicated well-known characteristics of the tests. The Student's t-test yielded conservative Type I error rates under population non-normality. For example, the Type I error rates for the exponential distribution for $n_1 = 5$, $n_2 = 15$ was 0.0276. Similarly, the result for the Chi-square distribution (df = 1) was 0.0180. However, the Type I error rates for all conditions studied for the Wilcoxon Rank Sum test and the permutation t-tests were within sampling error of nominal alpha.

Power Results

The comparative power results for the normal distribution also replicated well-known results in the literature. The t and the permutation t-tests' statistical power were nearly indistinguishable. The Wilcoxon Rank Sum test's power was either the same, or slightly less, as noted, for example, in Figure 1. As suggested by asymptotic theory, the maximum power advantage of the two t-tests over the Wilcoxon test was only about 0.04.

The results for the exponential distribution ($\mu = \sigma = 1$) with the different shifts in location, as reflected in Figure 2, demonstrates the Wilcoxon test is more powerful than the t and permutation t-tests, of which the latter two have essentially the same power. As shown in Figure 3, the power properties for the Chi-square distribution (df = 1) indicates the same power advantages for the Wilcoxon Rank-Sum test, with the t-test and

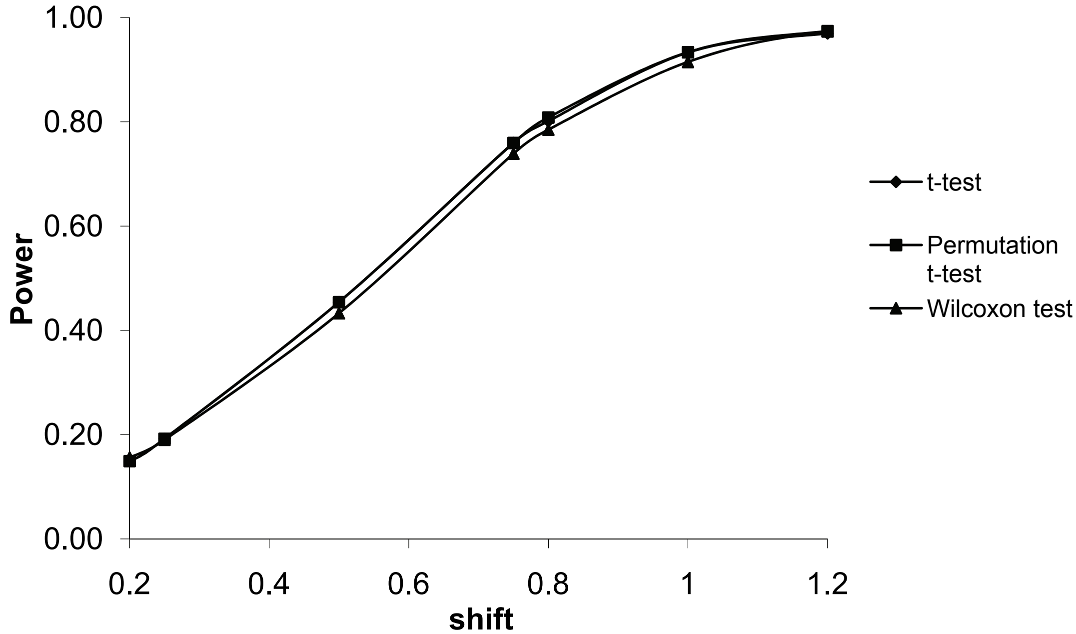

Figure 1: Shift vs. Power in the Normal Distribution for Sample Sizes $n_1 = n_2 = 20$

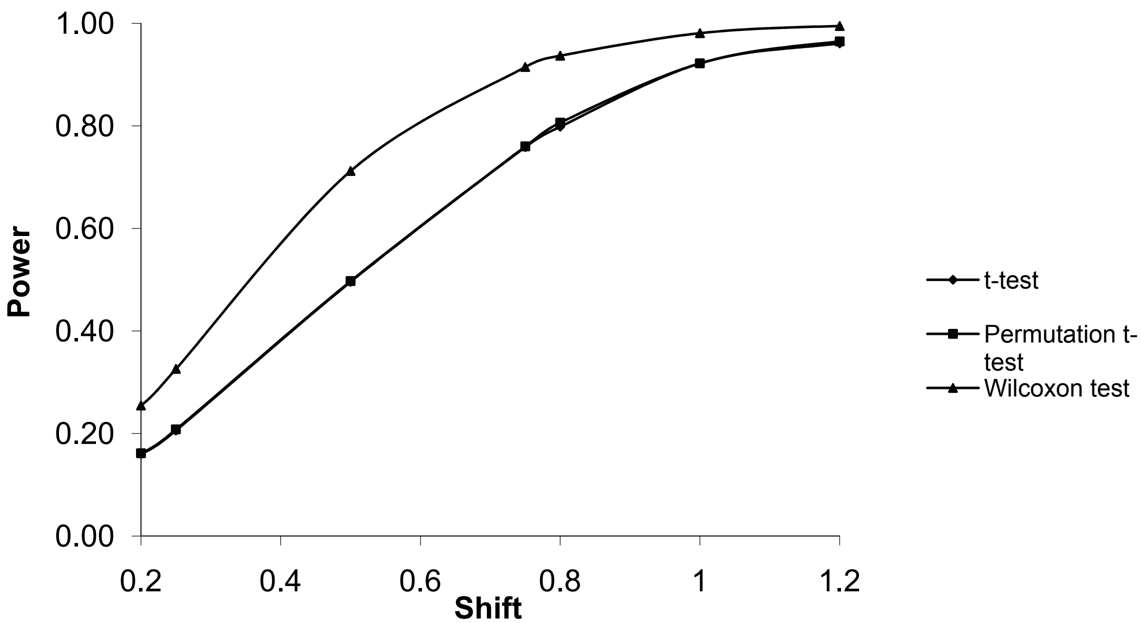

Figure 2: Shift vs. Power in the Exponential Distribution for Sample Sizes $n_1 = n_2 = 20$

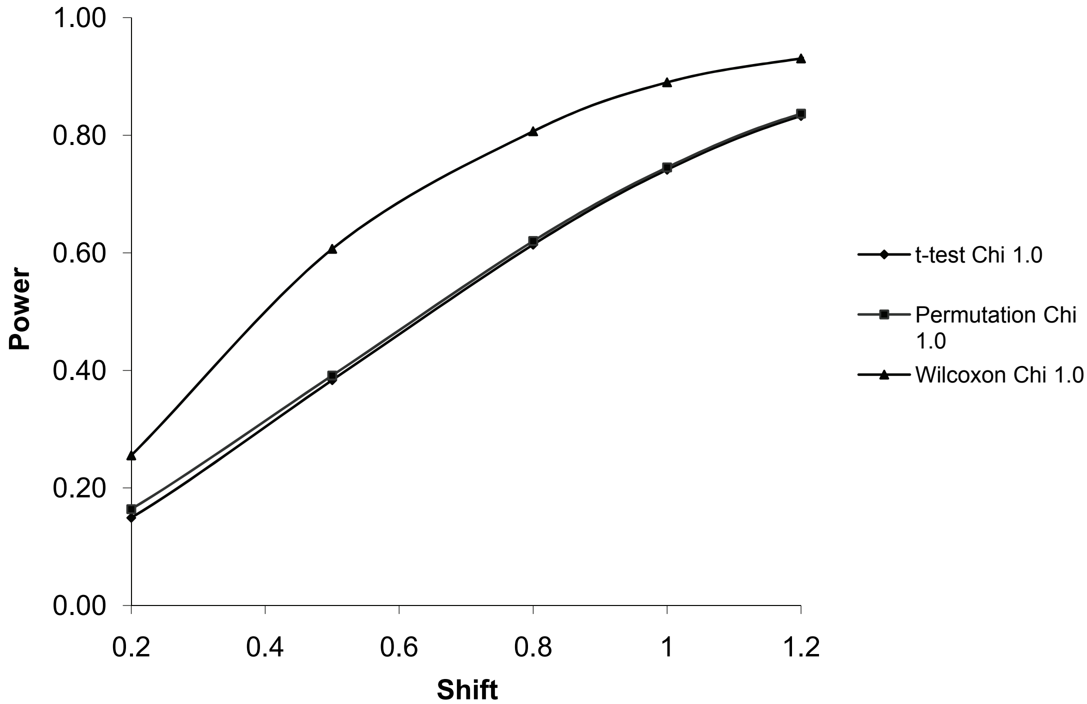

Figure 3: Shift vs. Power in the Chi-square Distribution (df = 1) for Sample Sizes $n_1 = n_2 = 10$

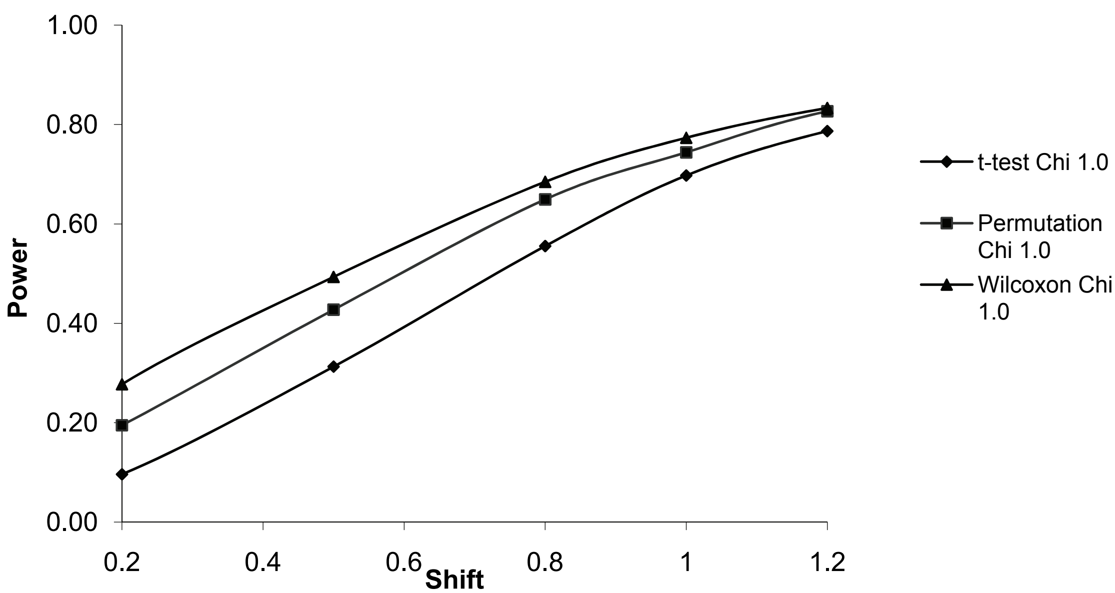

Figure 4: Shift vs. Power in the Chi-square Distribution (df = 1) for Sample Sizes $n_1 = 5$ & $n_2 = 15$

permutation t-test presenting nearly identical and substantially less statistical power. As indicated in Figure 4, the power results for the Chi-square distribution (df = 1) and unequal sample sizes indicated the permutation test became more competitive than the Student's t-test, but both tests remained considerably less powerful than the Wilcoxon Rank-Sum test.

Conclusion

Although Edgington (1995), Good (1994), and many others have presumed that the permutation t-test would be considerably more powerful than nonparametric tests, such as the Wilcoxon Rank-Sum test, the results of this Monte Carlo simulation did not support their opinion. These results pertain to the detection of a treatment modeled as a shift in location parameter, and of course, are based on the distributions, sample sizes, and the α level studied.

The primary answer provided by this simulation study is that the permutation test, in the context of the two independent samples layout, follows the depressed power spectrum of the Student's t-test, and not the superior spectrum afforded by the Wilcoxon test. Therefore, workers in applied research would be better served, when testing hypotheses of shift in location parameter, to use the nonparametric test instead of the permutation test.

Secondary results, interestingly, confirmed that the permutation t-test provides considerable power advantages over the Student's t-test for unbalanced sample sizes (e.g., Lu, Chase, & Li, 2001).

References

Adams, D. C. & Anthony, C. D. (1996). Using randomization techniques to analyse behavioural data. *Animal Behaviour*, *54*(4), 733-738.

Blair, R. C. (1985, March 31-April 4). Some comments on the statistical treatment of ranks. Paper presented at the 1985 AERA/NCME annual meeting, Chicago, IL.

Blair, R. C. & Higgins, J.J. (1980b). A comparison of the power of the Wilcoxon's rank-sum statistic to that of student's t statistic under various non-normal distributions. *Journal of Educational Statistics*, *5*, 309-335.

Blair, R. C. & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's sign-ranks test under various population shapes. *Psychological Bulletin*, *97*, 119-128.

Blair, R. C., Higgins, J.J. & Smitley, W.D. (1980). On the relative power of the U and t tests. *British Journal of Mathematical and Statistical Psychology*, *33*, 114-120.

Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, *40*, 26-42.

Bradley, J. V. (1968). *Distribution-Free Statistical Tests*. Englewood Cliffs, NJ: Prentice-Hall.

Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics*, *15*(2), 171-185.

Edgington, E. S. (1995). *Randomization Tests*. (3rd ed). New York, NY: Marcel Dekker.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York, NY: Springer-Verlag.

Good, P. (2002). Extentions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods*, *1*,(2), 243-247.

Higgins, J. J. & Blair, R. C. (2000, February). Letter to the Editor. *The American Statistician*, *54*, 86.

Hodges, J. & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t test. *Annals of Mathematical Statistics*, 27, 324-335.

Lehmann, E.L. & D'Abrera, H.J. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. New York, NY: McGraw-Hill.

Lu, M., Chase, G. & Li, S. (2001). Permutation tests and other tests statistics for ill-behaved data: Experience of the NINDS t-PA stroke trial. *Communications in Statistics-Theory and Methods*, *30*(7), 1481- 1496.

Ludbrook, J. & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, *52*(2), 127-133.

Maritz, J. S. (1981). *Distribution Free Methods*. London, England: Chapman and Hall.

Mielke, P. W. & Berry, K. J. (2001). *Permutation Methods: A Distance Function Approach*. New York, NY: Springer.

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

Sawilowsky, S. S. (1993). Comments on using alternatives to normal theory statistics in social and behavioral science. *Canadian Psychology*, *34*, 398-406.

Sawilowsky, S.S. & Blair, R.C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin*, *111*, 353-360.