Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 6

11-1-2004

Monte Carlo Evaluation of Ordinal d with Improved Confidence Interval

Du Feng
Texas Tech University, du.feng@ttu.edu

Norman Cliff
University of Southern California

DOI: 10.22237/jmasm/1099267560

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 322-332

1538 – 9472/04/$95.00

Monte Carlo Evaluation of Ordinal *d* with Improved Confidence Interval

Du Feng
Texas Tech University

Norman Cliff (Emeritus)
University of Southern California

This article reports a Monte Carlo evaluation of ordinal statistic *d* with modified confidence intervals (CI) for location comparison of two independent groups under various conditions. Type I error rate, power, and coverage of CI of *d* were compared to those of the Welch's *t*-test.

Key words: *d* statistic, Welch's *t*, computer simulation, Monte Carlo study

Introduction

One of the most commonly asked questions in social, behavioral, and biomedical research is concerned with whether scores from one group tend to be higher than those from the other (e.g., treatment effects). This type of location comparison questions (or two-sample problems) is usually answered by parametric tests such as Student's *t* test or Welch's *t* test, which requires interval level of measurement of the test variables. However, many behavioral and social variables have only ordinal justification (e.g., Likert-scaled data), thus, performing equivalence testing of two means can yield misleading results. Furthermore, Student's *t* test is known to be not robust when the normality and/or homogeneity of variance assumptions are violated (e.g., Wilcox, 1990, 1991), as they often are in empirical studies (Micceri, 1989; Wilcox, 1996, p.135). Although Welch's test was found to improve on *t* test under violations of these assumptions, ordinal methods are more appropriate, and can be more powerful, than the t tests for ordinal data.

Du Feng is an Associate Professor in the Department of Human Development and Family Studies, Box 41162, Lubbock, Texas 79409-1162. Email: du.feng@ttu.edu. Norman Cliff is a former president of the Psychometric Society, Society for Multivariate Experimental Psychology, and of Division 5 of the American Psychological Association. The authors wish to thank Dr. Rand Wilcox for his comments on an early draft of the manuscript.

Definition of *d*

Cliff (1993) introduced a dominance analysis summarized by the ordinal statistic *d*, which compares the proportion of times a score from one group or under one condition is higher than a score from the other, to the proportion of times when the reverse is true. The population analog of *d* is called δ (often written Δ). For random variables X_1 and X_2, $\delta = \Pr\{x_1 > x_2\} - \Pr\{x_1 < x_2\}$. It is equivalent to the form of Kendall's τ called Somer's *d* (Somer, 1968) when one variable is dichotomous. This measure was introduced and discussed by nonparametric statistics books for years (e.g., Agresti, 1984; Hettmansperger, 1984; Randles & Wolfe, 1979), and its application was emphasized and extended by Cliff (1991, 1993, 1996).

Advantages of the ordinal statistics over the classical ones have been suggested repeatedly, including their robustness and power under departure from normality or equal variance assumptions, being invariant under monotonic transformation, suitability for much behavioral data which can only be given ordinal-scale status, and their descriptive superiority (Caruso & Cliff, 1997; Cliff, 1993; Long, Feng, & Cliff, 2003). From its definition, we can see that δ is the effect size itself. It is more directly related to the research question often asked: whether scores in one group or under one condition tend to be higher than those in another, than is through some kind of comparison of means or medians. The sample *d* as defined is an unbiased estimate of δ

$$d = \frac{\#(x_i > x_j) - \#(x_i < x_j)}{n_1 n_2}, \quad (1)$$

where n_1 and n_2 are the sample sizes for x_i and x_j, respectively.

It was noted that δ is a simple transformation of a measure, $p = \Pr\{x_1 > x_2\}$, proposed by Birnbaum (1956): $p = (\delta + 1) / 2$, which is estimated by a "common language effect size statistic" (McGraw & Wong, 1992), when there is no ties between random scores from the two groups (Long, Feng, & Cliff, 2003). However, δ has advantages over *p* because it takes into account ties in the data (Long, Feng, & Cliff, 2003). Similarly, Vargha and Delaney (2000) proposed a generalization of the "common language effect size statistic" (CL) suggested by McGraw and Wong (1992), in order to take into account ties between the two groups scores. They called the generalization "*A* measure of stochastic superiority," which was defined as $A = \Pr\{x_1 > x_2\} + .5 \Pr\{x_1 = x_2\}$. It was noted that *A* is simply a linear transformation of δ: $A = (\delta + 1)/2$ (Vargha & Delaney, 2000, p.104).

Inferences About δ

With traditional ordinal methods, for example, the Wilcoxon-Mann-Whitney (WMW) rank-sum test (Mann & Whitney, 1947; Wilcoxon, 1945), inferences are usually based on a randomization hypothesis which assumes that the two populations are identically distributed under the null hypothesis. The identical distribution assumption makes the test tend to be sensitive to differences in spread (also called "scale") and shape of the two distributions. However, this assumption is not necessary for making inferences about δ, because the sampling distribution of *d* is asymptotically normal and normal-based inferences can be made, with σ_d^2 being estimated from the sample. Several researchers (Birnbaum 1956; Cliff, 1991, 1993, 1996; Fligner & Policello, 1981; Hettmansperger, 1984; Mee, 1990; Siegel & Castellan, 1988; Zaremba, 1962) have suggested ways of making inferences about δ based on *d* with the sample estimate of its variance, and described the calculation of the sample estimate of σ_d^2.

Fligner and Policello (1981) introduced a robust version of the WMW test for comparing the medians of two independent continuous distributions, and tested behavior of *d*, using the sample estimate of its variance. Their results indicated that *d* behaved well in small samples in terms of Type I error rate and power over a variety of conditions of population distribution. Cliff (1993) suggested a modification of Fligner and Policello's (1981) procedure by deriving an unbiased sample estimate of the variance of *d* and setting a minimum allowable value for it in order to increase the efficiency of the estimate and to eliminate impossible values. Defining a dominance variable, which represents the direction of differences between scores, as: $d_{ij} = \text{sign}(x_{i1} - x_{j2})$, where x_{i1} represents any observation in the first group, x_{j2} in the second, Cliff (1993) showed that variance of *d* can be expressed as

$$\sigma_d^2 = \frac{(n_1 - 1)\sigma_{d_{i.}}^2 + (n_2 - 1)\sigma_{d_{.j}}^2 + \sigma_{d_{ij}}^2}{n_1 n_2}, \quad (2)$$

where $d_{i.}$ is

$$d_{i.} = \frac{\#(x_i > x_j) - \#(x_i < x_j)}{n_1}, \quad (3)$$

and similarly for $d_{.j}$.

The unbiased sample estimate of σ_d^2 was shown to be

$$s_d^2 = \frac{n_1^2 \Sigma(d_{i.} - d)^2 + n_2^2 \Sigma(d_{.j} - d)^2 - \Sigma\Sigma(d_{ij} - d)^2}{n_1 n_2 (n_1 - 1)(n_2 - 1)}. \quad (4)$$

To eliminate possible negative estimate of variance, $(1 - d^2)/(n_1 n_2 - 1)$ was introduced by Cliff (1993, 1996) as the minimum allowable value for s_d^2. For detailed discussion on the σ_d^2 and its components, or the formulas presented above, see Cliff (1993, 1996).

Modification of CI for δ

The CI for δ is traditionally computed by $(d - z_{\alpha/2}\, s_d,\ d + z_{\alpha/2}\, s_d)$. However, this CI was

found to be unsatisfactory in Monte Carlo studies (Feng & Cliff, 1995; Vargha & Delaney, 2000). Delaney and Vargha (2002) used modifications of CI for δ that consisted of using Welch-like *df*s. They adopted these *df*s from Fligner and Policello (1981) procedure and Brunner-Munzel test (2000), and reported that these modifications improved performance of *d* (Delaney & Vargha, 2002).

These modifications, however, were used without paying attention to the specific situations in which *d* with traditional CI performed poorly. Long, Feng, and Cliff (2003) pointed out two reasons why *d* with traditional CI was unsatisfactory. One reason has to do with a zero estimated variance for *d* when $d = \pm 1$, in which case the conventional CI reduces to a point $\delta = \pm 1$. The other reason is that the traditional symmetric CI does not take into account the negative correlation between σ_d^2 and δ. They proposed using an asymmetric CI to account for boundary effects on the variance of *d* due to the negative correlation between σ_d^2 and δ. When $d \neq \pm 1$, using sample estimates of variance of *d*, the asymmetric CI for δ can be constructed based on the following equation:

$$\delta = \frac{d - d^3 \pm t_{\alpha/2} s_d (1 - 2d^2 + d^4 + t_{\alpha/2}^2 s_d^2)^{1/2}}{1 - d^2 + t_{\alpha/2}^2 s_d^2}. \quad (5)$$

When $d = \pm 1$, a conservative approach, leading to relatively wide CI, is to assume the maximum possible variance for *d*, given δ. The maximum possible variance ($\sigma_{d_m}^2$) occurs when the scores in one group are bimodal with all the scores in the other group falling between the modes, leading to a variance of

$$\sigma_{d_m}^2 = (1 - \delta^2)/n_b , \quad (6)$$

where n_b is the sample size of the bimodal group.

This relation between σ_d^2 and δ^2 in the extreme case was used in constructing a CI for δ when $d = \pm 1$. The method is similar to the one used in constructing a CI for population proportion from a sample proportion (see Hayes, 1973, p.379). Assuming that $(d - \delta)/\sigma_{d_m} \sim N(0,1)$, we have the CI with confidence level $1-\alpha$: $Z_{\alpha/2} < (d - \delta)/\sigma_{d_m} < Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the critical *z*-score at the selected α level. The upper and lower limits of the CI for δ are the solutions of the equation

$$Z_{\alpha/2}^2 = \frac{(d - \delta)^2}{\sigma_{d_m}^2}. \quad (7)$$

Inserting Equation (6) to the above for $\sigma_{d_m}^2$, when $d = 1$, the solution of Equation (7) gives

$$\delta = \frac{(n_b - Z_{\alpha/2}^2)}{(n_b + Z_{\alpha/2}^2)} \quad (8)$$

as the lower limit for the CI when $d = 1$ (in which case the upper limit is 1); and upper limit of the CI when $d = -1$ (in which case the lower limit is -1). With unequal groups, a conservative solution is to use the smaller sample size as n_b in Equation (8). This modification obviates the necessity of using a minimum allowable variance of *d*.

Methodology

A simulation study comparing rank *t* test, rank Welch test, Fligner-Policello test, and the *d* test found *d* to have inflated Type I error rate (Vargha & Delaney, 2000). However, the above mentioned modifications of CI was not used in this existing study. The primary purpose of the current study was to evaluate the performance of *d* with modifications of CI that were made based on theoretical and empirical concerns.

A Monte Carlo study was carried out in a variety of situations. To provide a basis for comparison for the behavior of *d*, the *t*-test with unpooled variance and Welch's adjustment of *df* (referred to as Welch's *t*, or t_w) was included in the analyses. Although this is known to be not completely robust (Wilcox, 1990), it is reasonably so for moderate variance heterogeneity, and it is clearly preferable to Student's *t*. It sacrifices a little power relative to the latter, but the sacrifice is realistic, especially in forming CI. It is now widely available in

statistical packages and is sometimes even the default statistic for mean comparisons.

Samples of small (n = 10) to moderate (n = 30) sizes were taken repeatedly from a large number of pairs of uncorrelated populations. In simulating the data, five factors were manipulated: form, mean, variance, skewness of the parent distributions, and sample size. Then, statistical inferences about δ were computed based on each selected pair of samples, and two-sided d and t_w tests at the .05 significance level were performed to compare the two independent groups. Subroutines of IMSL library were called by Fortran programs to generate the populations and samples. Another Fortran program was written to compute statistical inferences about δ for two independent groups and to perform d and t_w tests.

The intention of the present study was to investigate a variety of situations so that the results could be generalized to a wide spectrum of behavioral data. Behavioral variables are often strongly skewed (Miccerri, 1989; Wilcox, 1990, 1991), with concomitant kurtosis, whereas thick-tailed, but symmetric, distributions seem less common. Variables are often bounded by zero, and many are bounded at both ends. Furthermore, distributions differing in location can also differ in scale and/or skewness. Therefore, four families of distributions were selected for the Monte Carlo study: normal, skewed (defined below), chi-square, and beta-distributions. Chi-square and beta-distributions were employed to simulate one-side-bounded and two-side-bounded data with various degrees of skewness, respectively.

Within each family of distributions, certain combinations of means and variances were selected so that δ ranged from .3 to .8. The selection of effect sizes, in terms of δ, conforms to Cohen's (1988) guidelines for small, medium, and large effects for comparable location models.

Normal Distribution

The normal distributions selected had μ of 0, 1, 2, or 3, and σ^2 of 1, 4, or 9. While all pairs of groups with these means and variances were considered, only a subset of them, representing typical results, are reported here. With symmetric distributions, the null hypothesis for the d analyses, $H_0 : \delta = 0$, is true when the null hypothesis for t_w, $H_0 : \mu_1 = \mu_2$, is true.

Skewed Distribution

Although there appears to be no satisfactory guidelines on what values of skewness are realistic, some studies found that estimated skewness of 2 was not uncommon (Micceri, 1989; Wilcox, 1990). Thus, skewnesses of -2, 0, and 2 were used to examine the effect of unequal skewnesses. The logistic inverse transformation (Ramberg et al., 1979): $\pm \log (U - 1)^{-1}$, where U represents a uniform distribution on the interval zero to one ($0 \leq U \leq 1$), was used to generate skewed data. This yields distributions having skewnesses of 2 or -2. The transformed data were then re-scaled to have μ of -3, -1, 0, 1, 2, or 3, with σ^2 of 1, 4, or 9.

These skewed distributions also have heavier tails than the normal distribution; their kurtosis tends to be around 5. To avoid a possible effect of unequal kurtosis, and separate it from the effect of unequal skewness, the h-transformation: $Ze^{hZ^2/2}$ (Hoaglin, 1985), where Z is N(0,1), was applied to generate symmetric populations with greater kurtosis. $h \approx .126$ results in kurtosis of around 5, which is comparable to kurtosis of the skewed distributions.

Given the levels of mean, variance, and skewness, there can be 54 different kinds of combinations for each group, and the number is squared when two groups are involved. However, only some representative combinations were selected, and a subset of these are reported here. Unlike in the normal case, for skewed data, the null hypothesis regarding δ and the null hypothesis regarding ($\mu_1 - \mu_2$) are not necessarily both true or both false, although effects are quite small. Cases when both H_0's are true or false, as well as one of them is true while the other is false, were included.

Chi-square Distribution

The one-side-bounded data were simulated using chi-square distributions with df ranging from 2 to 32. Certain combinations of the population groups were selected so that the

effect size, δ, fell into the low (.3) to high (.8) range. Several chi-square variates were rescaled by multiplying by constants in order to obtain the desired effect sizes.

Beta Distribution

The two-side-bounded data were generated, according to beta distributions with the first parameter (p) and the second parameter (q) ranging from 1 to 14. Again, certain population groups were selected for comparison, so that δ ranged from .3 to .8.

The null cases for bounded data were those when the two groups had identical chi-square or beta distributions. For the non-null cases, again, the populations compared could have equal or unequal variance, skewness, and kurtosis. For non-normal data, four non-null situations were considered: when two groups were (a) the same in shape (skewness and kurtosis) and scale (variance); (b) the same in shape but different in scale; (c) the same in scale but different in shape; and (d) different in shape and scale.

Sample size, particularly differences in sample size, can profoundly affect the behavior of location comparisons. For each population, observations were simulated for two independent groups using four combinations of the sample sizes n_1 = 10, 30, and n_2 = 10, 30. Both d and t_w tests were performed for the same data at the α = .05 significance level. Two thousand simulation replications were employed under each distributional situation, so that for nominal α = .05 and the 95% CI, a .01 difference is significant.

For example, empirical α's that are higher than .06 are considered significantly higher than the nominal level .05; similarly, CI coverages that are lower than .94 are considered significantly lower than the nominal .95. With 2000 replications and α = .05 for the proportions test, the power of the test to detect a departure of $\alpha \pm 1/2\alpha$, which was defined as the "liberal" tolerance criterion (Bradley, 1978) for robustness of Monte Carlo experiments, is .996; the power to detect a departure of $\alpha \pm 1/4\alpha$, the "intermediate" criterion (Robey & Barcikowski, 1992), is .7 (Cohen, 1988; Robey & Barcikowski, 1992).

The d and t_w tests were evaluated and compared in terms of three criteria: empirical Type I error rate, power, and CI coverage. The three criteria evaluate the tests from three different aspects. Coverage of CI has not been addressed as much as the other two by similar studies, though it is equally important and informative, and it is not necessarily implied by the others.

The proportion of the 2000 statistics that exceed the appropriate .05 critical values in the null case is the empirical Type I error rate. It is an estimate of the actual probability of a Type I error. Power is estimated by the proportion of rejection in the right direction at the .05 level in non-null cases. The CI coverage probability is estimated by the proportion of times that the CI constructed by each method covers the corresponding population parameter.

Results

Comparison of empirical α of d and t_w, revealed that with the adjusted CI, d gave rejection rates that were at or below .05 under all circumstances, tending to be conservative when at least one group was small (n = 10). On the other hand, use of the simple traditional CI led to liberal empirical α's (greater than .06) when at least one group was small, particularly when the small n was paired with a larger variance. Welch's t gave several α's above .06 when group sizes were unequal. It should be noted that none of these departures were above the liberal criterion, even though the range of conditions studied was wide.

The findings about the performance of d are similar to those of Fligner and Policello's (1981) in that d behaved well in terms of controlling the probability of Type I errors, but d appeared to be more conservative in this study with the adjustments on the CI for δ. Table 1 summarizes empirical Type I error rates of d and t_w.

Table 1. Empirical Type I Error Rate of d and t_W for $\alpha = .05$

			n_1=10,n_2=10		n_1=30,n_2=30		n_1=10,n_2=30		n_1=30,n_2=10	
	σ_1:σ_2	γ_1-γ_2	d	t_W	d	t_W	d	t_W	d	t_W
Normal	1:1	0	.021$^-$	.035$^-$	.048	.052	.041	.051	.037$^-$	.047
	1:3	0	.031$^-$	.056	.039$^-$	.048	.048	.057	.032$^-$	.048
Skewed	1:1	0	.038$^-$	.044	.047	.048	.038$^-$	.064$^+$	.041	.064$^+$
	1:3	4	.029$^-$	--*	.049	--*	.050	--*	.032$^-$	--*
Chi-square	1:1	0	.028$^-$	.042	.050	.057	.039$^-$	.060$^+$	.049	.061$^+$
Beta	1:1	0	.033$^-$	.050	.048	0.52	.037$^-$	.053	.039$^-$	.057

$^+$ At least two standard deviation above .05, computed as if $\alpha = .05$.

$^-$ At least two standard deviation below .05, computed as if $\alpha = .05$.

* No Type I error rate of t_W reported because this is a non-null case for means.

Power

Detailed results on the empirical power of the tests are summarized in Table 2. In general, t_w showed slightly higher power than d (when the adjusted CI was employed) in small samples. When both sample sizes were as larger as 30, d and t_w had similar power. However, it should be noted that a direct power comparison between the two statistics is not always valid, because they usually had different actual α level and different CI coverage as well. It is also noted that many of the conditions under which t_w had the power advantage are those where its Type I error rate was too high in the null case, or the CI coverage was inadequate. Thus its advantages are largely spurious.

The power of both tests increased with sample size, and with effect size, in the expected ways. However, it appeared that the sample size had a stronger effect on d than on t_w, given that with moderate samples ($n_1 = n_2 = 30$), the power advantage of t_w became less obvious or disappeared--d sometimes had slightly higher power than t_w. Figure 1 shows an example of this condition with chi-square distributions.

Power of d with unadjusted CI was slightly higher compared to the reported power with the adjustments. However, as noted, this slight gain in power is associated with higher Type I error rate and poorer CI coverage.

Coverage of CI

With the aforementioned adjustments on the CI for δ (i.e., the adjustment when $d = \pm 1$, and the asymmetric adjustment), d performed well in general in terms of CI coverage, with a few exceptions. This coverage appeared to be a negative function of δ. It was at or above the nominal $1 - \alpha$ level independent of sample size, the form of the population distributions, and

Table 2. Empirical Power of d and tw for * = .05

			n_1=10,n_2=10		n_1=30,n_2=30		n_1=10,n_2=30		n_1=30,n_2=10	
δ	σ_1:σ_2	sk_1-sk_2	*d*	t_w	*d*	t_w	*d*	t_w	*d*	t_w
Normal Distribution										
.218	3:2	0.0	.093	.132	.269	.299	.120	.157	.177	.207
.363	1:1	0.0	.212	.275	.674	.709	.338	.398	.360	.425
.473	3:1	0.0	.280	.438	.848	.918	.301	.442	.773	.869
.520	1:1	0.0	.464	.553	.961	.972	.661	.719	.659	.705
.711	1:1	0.0	.803	.868	1.0	1.0	.947	.967	.943	.971
.820	2:1	0.0	.910	.967	1.0	1.0	.927	.984	1.0	1.0
Skewed Distribution										
.227	1:3	-2.0	.078	--*	.249	--*	.208	--*	.092	--*
.397	1:1	0.0	.254	.218	.743	.513	.158	.037	.311	.368
.472	1:1	4.0	.402	.566	.899	.997	.578	.852	.588	.844
.503	3:1	0.0	.422	.467	.940	.981	.216	.070	.731	.898
.781	1:3	-4.0	.905	.939	1.0	1.0	1.0	1.0	.971	.956
.816	3:1	0.0	.948	.907	1.0	1.0	.999	.941	.992	.992
Chi-square Distribution										
.242	14:1	.9	.089	.191	.289	.712	.096	.186	.259	.661
.346	1:1	-.5	.206	.218	.632	.524	.356	.225	.312	.381
.498	2:1	-.6	.405	.407	.942	.887	.642	.465	.615	.675
.662	1:1	-.1	.728	.794	.998	.998	.941	.942	.845	.915
.807	5:1	-.1	.939	.971	1.0	1.0	.976	.982	.998	.999
.835	1:1	-1.1	.945	.966	1.0	1.0	1.0	1.0	.962	.982
Beta Distribution										
.291	4.5:1	-.1	.130	.249	.463	.768	.143	.236	.334	.611
.327	2.5:1	-.4	.181	.250	.585	.723	.217	.241	.321	.481
.411	1:1	-.1	.276	.343	.802	.837	.440	.507	.454	.512
.553	7:1	-.5	.481	.612	.968	.994	.580	.650	.840	.940
.650	1:1	-.9	.700	.774	.998	.997	.941	.934	.818	.879
.814	12:1	-1.6	.904	.978	1.0	1.0	.937	.991	1.0	1.0

* No power of t_w reported because these are null cases for means.

Figure 1. Empirical power curve with chi-square distribution when $n_1 = n_2 = 30$.

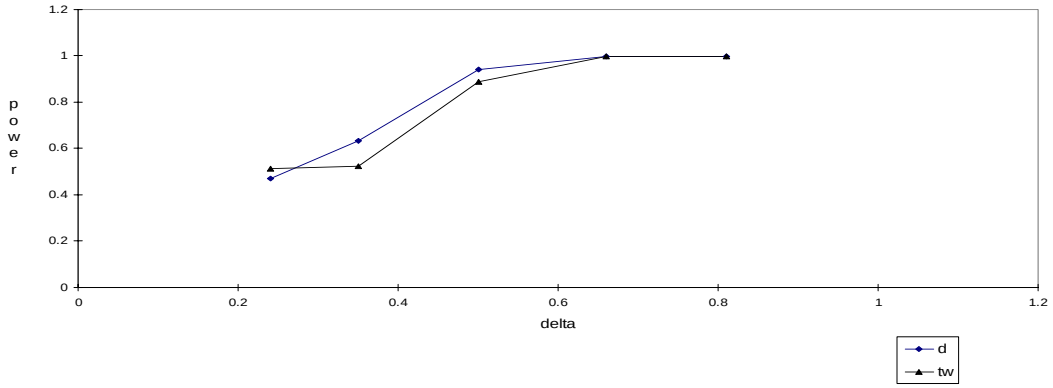

Figure 1
Empirical Power Curve When n1 = n2 = 30
- Chi-Square Distribution

variance ratio, skewness, and boundedness of the populations compared, unless when δ was quite high (above .7). But it rarely dropped below .93 under all conditions considered. The adjustments provided improvement over the unadjusted CI--the coverage was lower without the adjustments when δ was above .7.

The Welch's *t*-test yielded good CI coverage for μ_d with normal data, regardless of variance ratio and sample size. However, it was not robust to skewness and nonnormality. The coverage was particularly poor when skewness was combined with heterogeneity of variance, or when high population variance ratio was combined with boundedness and/or small or unequal sample sizes. Table 3 shows results on the empirical CI coverage of *d* and t_w.

Conclusion

The ordinal method *d* does not involve excessive elaboration and complicated statistical analyses. Its concept can be easily understood by nonstatisticians. The aforementioned computer program for independent groups *d* analysis is easy to implement. Its output provides descriptive information, not only the null hypothesis is tested, but also a CI is provided. In addition, a dominance matrix that the program produces is a useful visual aid to the test.

It was a preliminary purpose of this study to evaluate the performance of *d* with comparison to the Welch's *t*. The performance of *d* was evaluated in terms of Type I error rate, power, and CI coverage using a variety of normal and nonnormal data, and was compared to that of Welch's *t*-test. The findings based on simulations generally show that *d*, with adjusted CI, has good control over α under all conditions considered. Welch's *t* controls α at its nominal level with normal data, but sometimes fails to do so under nonnormality. Theory indicates that unequal sample sizes and unequal skewnesses would affect the robustness of t_w (Wilcox, 1990), and the results support this conclusion.

The results on t_w is also consistent with previous researches which found the t_w to be robust when $n_1 = n_2$ (Tan, 1982; Wilcox, 1990), and which showed that t_w was not robust in terms of Type I errors when the two groups had unequal variances, unequal sample sizes, and unequal skewnesses (Wilcox, 1990). Although, t_w behaved better here than Wilcox (1990) reported, probably because the levels of nonnormality examined in this study were not as high as in Wilcox (1990).

Adjustments of the CI for δ were proposed here, and it was examined whether and to what extent the adjustments improved the distributional behavior of *d*. The simulation results suggest that these adjustments improve the performance of *d* in term of Type I error rate and coverage, with a slight loss of power. However, the coverage is not completely satisfactory—it is adequate when δ is not too high, but can be low when the population δ is close to 1. Perhaps even further modification on the construction of the CI for δ is needed.

For both *d* and t_w, using normal or nonnormal data, under each selected effect size, the performance of the tests were better when the sample sizes were larger. This is accounted for by the central limit effect.

The findings of this study are partly consistent with those of Fligner and Policello's (1981) in that both studies suggest that the small sample behavior of *d* is good in terms of Type I error rate under normality, and it is robust when there is shift in scale. However, in our study, without the adjustments on the CI for δ, *d* sometimes appears to be more liberal in terms of actual α.

In this article, skewed, chi-square, and beta-distributions were selected for the purpose of assessment. More types of nonnormal distributions, such as heavy-tailed distributions, can be used in future simulation studies testing the behavior of the statistics. Further more, the distribution characteristics of the *d* statistic, its variance, and other components such as $d_{i.}$ and $d_{.j}$ should be investigated in further detail.

Several ad hoc analyses of *d* and s_d were carried out in an attempt to shed light on the reasons both for its good behavior and for the exceptions. No conclusions are possible so far, but some directions for investigation are suggested by these analyses. One aspect of the regular behavior of *d* may lie in the relative stability of s_d^2 as an estimate of σ_d^2 under most circumstances.

Table 3. Estimated Confidence Interval Coverage Probability of d and t_W for $\alpha = .05$

			n_1=10,n_2=10		n_1=30,n_2=30		n_1=10,n_2=30		n_1=30,n_2=10	
δ	σ_1:σ_2	sk_1-sk_2	d	t_W	d	t_W	d	t_W	d	t_W
Normal Distribution										
.218	3:2	0.0	.962$^+$	.950	.955	.951	.959	.945	.957	.945
.363	1:1	0.0	.964$^+$	.949	.952	.946	.953	.943	.954	.947
.473	3:1	0.0	.951	.946	.959	.950	.942	.946	.950	.945
.520	1:1	0.0	.951	.958	.960$^+$	.945	.956	.950	.942	.949
.711	1:1	0.0	.933$^-$	.945	.951	.956	.949	.953	.945	.942
.820	2:1	0.0	.921$^-$	.950	.930$^-$	.955	.862$^-$	.941	.946	.955
Skewed Distribution										
.227	1:3	-2.0	.964$^+$	.918$^-$	.963$^+$	.935$^-$	.954	.941	.971$^+$	.896$^-$
.397	1:1	0.0	.962$^+$	.965$^+$	.960$^+$	.950	.956	.913$^-$	.942	.937$^-$
.472	1:1	4.0	.958	.915$^-$	.957	.951	.961$^+$	.925$^-$	.963$^+$	.926$^-$
.503	3:1	0.0	.956	.913$^-$	.964$^+$	.932$^-$	.966$^+$	.851$^-$	.949	.948
.781	1:3	-4.0	.961$^+$	.914$^-$	.950	.924$^-$	.953	.938$^-$	.950	.910$^-$
.816	3:1	0.0	.938$^-$	.903$^-$	.944	.934$^-$	.951	.912$^-$	.908$^-$	.951
Chi-square Distribution										
.242	14:1	.9	.968$^+$	.916$^-$	.960$^+$	.934$^-$	.967$^+$	.923$^-$	.963$^+$	.942
.346	1:1	-.5	.958	.956	.951	.947	.962$^+$	.942	.950	.938$^-$
.498	2:1	-.6	.956	.956	.963$^+$	.951	.961$^+$	.931$^-$	.951	.953
.662	1:1	-.1	.947	.951	.947	.945	.951	.948	.928$^-$	.945
.807	5:1	-.1	.938$^-$	.951	.938$^-$	.944	.922$^-$	.939$^-$	.931$^-$	.957
.835	1:1	-1.1	.925$^-$	.951	.924$^-$	.954	.948	.956	.879$^-$	.943
Beta Distribution										
.291	4.5:1	-.1	.959	.944	.957	.949	.962$^+$	.925$^-$	.956	.950
.327	2.5:1	-.4	.965$^+$	.944	.953	.946	.962$^+$	.933$^-$	.958	.946
.411	1:1	-.1	.963$^+$	.955	.949	.948	.957	.950	.950	.946
.553	7:1	-.5	.954	.939$^-$	.955	.943	.946	.930$^-$	.951	.945
.650	1:1	-.9	.946	.959	.949	.949	.957	.950	.930$^-$	.935$^-$
.814	12:1	-1.6	.932$^-$	.934$^-$	.931$^-$	.945	.899$^-$	.941	.946	.954

$^+$ At least two standard deviation above .95, computed as if $\alpha = .05$.

$^-$ At least two standard deviation below .95, computed as if $\alpha = .05$.

The cases of relatively poor behavior may result from two sources. One source is the correlation between s_d and d that becomes quite strong when δ is fairly high. The asymmetric CI, given by Equation (10), is one attempt at compensating for this effect, but it seems not to be strong enough when δ is very high, and may be too strong when it is low, at the expense of power. It may also be that there are a few circumstances where s_d^2 is less well behaved, although we do not clearly understand what these circumstances are.

Understanding the behavior of d may be facilitated by noting that it, too, is a mean difference. Let (p_{11}, p_{12},…, p_{1n1}) be the values of a variable representing the proportion of x_{j2} scores that are less than each x_{i1}, respectively, and correspondingly for the second sample. That is, $p_{i1} = ½(d_{i.} + 1)$, and $p_{j2} = ½(d_{.j} + 1)$. Then d is the difference between the mean p_{i1} and the mean p_{j2}. Each p_{i1} reflects—although it does not equal—a corresponding value of a random variable P_1. Given a distribution $F_1(X_1)$ and correspondingly $F_2(X_2)$, then for any x_{i1}, $p_{i1} = F_2(x_{i1})$, and vice versa, and each has a distribution, $G_1(P_1)$ and $G_2(P_2)$, respectively. Therefore, the behavior of d depends on the nature of these distributions in much the same way that the behavior of the sample mean difference depends on F_1 and F_2. A difference is that p_{i1} is a binomial distribution of p_{i1} whose value depends on which x_{j2} happens to be in the sample. The two parts of the expression for the variance of d reflect these two aspects of the sampling process.

Not only does the variance of d depend on the variance of p_{ij}, but the other moments of its distribution depend on the other moments of their distributions. Thus, d is not distribution-free except in the limiting case where F_1 and F_2 coincide, but it depends on the distributions G_1 and G_2 rather than on F_1 and F_2. The fact that it behaves more robustly than t_w simply reflects the fact that the distributions that determine its behavior tend to have better properties that the distributions of the variables themselves. However, we should not be surprised if situations can be found where the opposite is true. These issues can be investigated in future studies.

In sum, this article has shown that d behaves quite well in small and moderate samples in terms of Type I error rate, power, and coverage of the CI, but not perfectly. The adjustments to the CI improved matters in terms of Type I error rate and coverage. This ordinal statistic is robust to nonnormality, heterogeneity of variance, and unequal sample sizes. Yet, there are a few exceptions to the good behavior of d, and further modification may be needed when the population δ is very close to 1 or -1.

Welch's *t*-test performs well under normality, but is not robust to nonnormality. Its Type I error rate is inflated, power is lowered, and coverage is inadequate when the populations are skewed, and when nonnormality is combined with unequal variances and/or unequal sample sizes. It is particularly sensitive to skewness.

The d has attractive characteristics as a description of location difference. It is a direct numerical reflection of the tendency for scores in one group to lie generally above those of another. It is also invariant under monotonic scale transformations, so conclusions about location need less qualification. The additional fact that its sampling behavior has to be rated as very good seems to lead to a conclusion that it is the method of choice for location comparison in many situations.

References

Agresti, A. (1984). Analysis of ordinal categorical data. New York: Wiley.

Best, D. J., & Rayner, J. S. W. (1987). Welch's approximate solution for the Behrens-Fisher problem. *Technometrics*, *29*, 205-210.

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

Caruso, J. C., & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for Spearman's rho. *Educational and Psychological Measurement*, *57*, 637-654.

Cliff, N. (1991). Ordinal methods in the study of change. In Collins, L. M. & Horn, J. (Eds.) Best methods for the analysis of change. Washington, D.C.: American Psychological Association, 34-46.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*, 494-509.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. New Jersey: Lawrence Erlbaum.

Cochran, W. G. (1951). Testing a linear relation among variances. *Biometrics*, *7*, 17-32.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, (2nd ed). Hillsdale, NJ: Erlbaum.

Davenport, J. M., & Webster, J. T. (1975). The Behrens-Fisher problem, an old solution revised. *Metrika*, *22*, 47-54.

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods*, *7*, 485-503.

Fligner, M. A. (1979). A class of distribution-free tests for the two-sample scale problem. *Journal of the American Statistical Association*, *74*, 889-893.

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

Hayes, W. L. (1973). *Statistics for the Social sciences*, (2nd Ed.). New York:Holt, Winehart, and Winston.

Hettmansperger, T. P. (1984). *Statistical inferences based on ranks*. New York: Wiley.

Hoaglin, D. C. (1985). Summarizing shape numerically: the g&h distributions. In Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.) *Exploring data tables, trends, and shapes*. New York: Wiley, 461-513

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50-60.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361-365.

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

O'Brien, P. C. (1988). Comparing two samples: Extension of the t, rank-sum, and log-rank tests. *Journal of the American Statistical Association*, *83*, 52-61.

Ramberg, J. S., Tadikamalla, P.R., Dudewicz, E.J., & Mykytka, E.F., (1979). A probability distribution and its uses in fitting data. *Technometrics*, *21*, 201-214.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. New York: Wiley.

Robey. R. R., & Barcikowski, R. S. (1992). Type I Error and the number of iteration in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283-288.

Tan, W. Y. (1982). Sampling distribution and robustness of t, F and variance ratio of two samples and ANOVA models with respect to departure from normality. *Communications in Statistics: Theory and Computation*, *11*, 2485-2511.

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

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-362.

Wilcox, R. R. (1990). Comparing the means of independent groups. *Biometric Journal*, *32*, 771-780.

Wilcox, R. R. (1991). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? Current *Directions in Psychological Science*, 1, 101-105.

Wilcoxon, F (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80-83.

Zaremba, S. K. (1962). A generalization of Wilcoxon's test. *Monatshefte fur Mathematik*, *66*, 359-370.