Psicológica (2012), 33, 391-418.

Correcting Two-Sample *z* and *t* Tests for Correlation: An Alternative to One-Sample Tests on Difference Scores

Donald W. Zimmerman*

Carleton University, Canada

In order to circumvent the influence of correlation in paired-samples and repeated measures experimental designs, researchers typically perform a one-sample Student *t* test on difference scores. That procedure entails some loss of power, because it employs $N - 1$ degrees of freedom instead of the $2N - 2$ degrees of freedom of the independent-samples *t* test. In the case of non-normal distributions, researchers typically substitute the Wilcoxon signed-ranks test for the one-sample *t* test. The present study explored an alternate strategy, using a modified two-sample *t* test with a correction for correlation, analogous to the "z test for correlated samples" used at one time for paired observations. For non-normal distributions, the same modified *t* test was performed on rank-transformed data. Simulations disclosed that this procedure protects the Type I error rate for moderate and large sample sizes, maintains power for normal distributions and several symmetric non-normal distributions, and substantially increases power for various skewed non-normal distributions.

Statistical analysis of paired-samples or repeated-measures experimental designs typically employs the one-sample Student *t* test on difference scores in place of the independent-samples *t* test. This method, widely used in the past, entails some loss of power, because the test on differences is necessarily based on $N - 1$ instead of $2N - 2$ degrees of freedom. In the first part of the last century, data from paired-samples was often analyzed in a different way. Many introductory textbooks in that period, focusing mainly on large-sample studies for which the z-test is appropriate, presented methods of analyzing what were called *correlated samples*, using a modification of the familiar two-sample *z* test. These

* Donald W. Zimmerman, Ph.D. Professor Emeritus, Carleton University. Ottawa. Canada. Mailing address: 1978 134A Street. Surrey, BC, V4A 6B6 Canada. Email: dwzimm@telus.net

methods calculated the standard deviation of a difference between means by the formula

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} - 2\rho_{12}\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}}, \tag{1}$$

where ρ_{12} is the correlation between X_1 and X_2, and based the standard error on this value when calculating the z statistic (see, example, Guilford & Fruchter, 1973; p. 154; Hays, 1988, pp. 313-315; McNemar, 1955, p. 85; Snedecor & Cochran, 1989, pp. 99-100). Recent textbooks sometimes include these formulas, although authors usually recommend the paired-samples t test, not the z test for correlated samples, as a practical method. Furthermore, after nonparametric methods became widely used to overcome non-normality, the Wilcoxon signed-ranks test typically was used in place of the t test on difference scores when the normality assumption was questionable.

The present study re-examined some two-sample significance tests based on paired data, using formulas containing correlation coefficients. Instead of the large-sample z test, however, it employed a version of the two-sample Student t test modified to allow for correlation (Zimmerman, Williams, & Zumbo, 1993; Zimmerman, 1997). Thus, tests were based on $2N - 2$ degrees of freedom, instead of the $N - 1$ degrees of freedom of the one-sample t test. And in the case of non-normal distributions, it employed the same modified two-sample t test on *rank-transformed* data, instead of the Wilcoxon signed-ranks test based on differences. For a variety of both normal and non-normal distributions, this strategy brought about some improvement in control of Type I error rates, as well as an increase in the power to detect differences.

A TWO-SAMPLE *T* TEST WITH A CORRECTION FOR CORRELATION

It is possible to derive a "t test for correlated samples" analogous to the "z test for correlated samples." An estimate of the population variance from sample data is based on the standard deviation of a difference between means given by equation (1), instead of the well known formula for the standard error of a difference (Zimmerman, Williams, & Zumbo, 1993). Because $N_1 = N_2 = N$, where N is the number of pairs in the paired-samples procedure, together with the conventional assumption $\sigma_1 = \sigma_2 = \sigma$, we obtain $\sqrt{2\sigma^2(1-\rho)/N}$ for the standard error of a difference between means. The weighted estimate of the population standard deviation from the two sample variances for this case is $\sqrt{[\sum(X_1-\bar{X}_1)^2+\sum(X_2-X_2)^2]/2(N-1)}$. Substituting this result for the above standard error σ in the usual expression for *t*, with equal sample sizes, gives the result

$$t' = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\left(\frac{\sum(X_1-\bar{X}_1)^2+\sum(X_2-\bar{X}_2)^2}{N(N-1)}\right)(1-\rho)}}, \tag{2}$$

where ρ is the population correlation. Further simplified, if *t* is the usual Student *t* statistic based on two independent samples of *N* observations each, then

$$t' = t/\sqrt{1-\rho}. \tag{3}$$

Because ρ is a constant, (3) indicates that a correlation between pairs increases or decreases the variance of t depending on whether ρ is positive or negative. In practice, with only sample values available, parameters of the distribution of t' are unknown. Sample distributions of the t and t' statistics based on 20,000 samples are shown in Figure 8 to be discussed below. The figure shows the reduction of the variance of the sample distribution resulting from correlation and the increase in variance after the correction.

In most research studies that analyze paired data, the value of the population correlation ρ is not known. In order to make use of equation (2) in practical significance testing, it is necessary to substitute a correlation coefficient estimated from sample data for the unknown population correlation. The present simulation study investigated how this modified test compares to the paired-samples *t* test with regard to Type I error rates and power, and how substitution of a sample estimate, *r*, for the population correlation, ρ, in equation (2) affects the accuracy of the result. The study also examined an analogous procedure in which the same test, performed on rank-transformed data replacing the original scores, is substituted for the Wilcoxon signed-ranks test.

Table 1 is a 2 x 6 classification of some two-sample tests of differences in location, based on, first, whether the population variance is known or estimated and, second, whether the population correlation is known or estimated. In the case in which the correlation is known, a further dichotomy is based on whether that correlation is zero or nonzero. The table shows the significance tests usually recommended in introductory textbooks for these various possibilities. The upper section is relevant to normal populations and the lower section to non-normal populations, where nonparametric methods are conventional. The present paper investigates the possibility of substituting a new two-sample test, analogous to the *z* test for correlated samples, for the one-sample tests on differences in both the upper and lower sections.

Table 1. Classification of significance tests considered appropriate for paired data with known and estimated population variances and correlation coefficients.

normal populations (parametric case)

	ρ known		ρ estimated
	$\rho = 0$	$\rho \neq 0$	
σ^2 known	*z* test	*z* test for correlated samples	*z* test for correlated samples
σ^2 estimated	*t* test	paired-samples *t* test	paired-samples *t* test

non-normal populations (nonparametric case)

	ρ known		ρ estimated
	$\rho = 0$	$\rho \neq 0$	
σ^2 known	Wilcoxon-Mann-Whitney test	Wilcoxon signed-ranks test	Wilcoxon signed-ranks test
σ^2 estimated	Wilcoxon-Mann-Whitney test	Wilcoxon signed-ranks test	Wilcoxon signed-ranks test

METHOD[1]

The study compared Type I error rates and power of several significance tests performed on correlated samples from normal and 9 non-normal distributions. Three of the non-normal distributions were symmetric and 6 were skewed. The significance tests were (1) the independent-samples Student *t* test, (2) the paired-samples Student *t* test, (3) the Wilcoxon-Mann-Whitney test, which assumes identical distribution functions irrespective of shape, and which is equivalent to the Student *t* test performed on rank-transformed data, (4) the Wilcoxon signed-ranks test, and (5) the modified *t* test described above, using sample correlation coefficients, *r*, in place of the population correlation ρ.

Simulations were performed using *Mathematica,* version 4.1, together with *Mathematica* statistical add-on packages. All sample values were transformed to have mean 0 and standard deviation 1. Constants were added to the scores in one sample in increments of fixed proportions of a standard deviation, in order to produce systematic differences in means and determine the power of the tests.

The correlation between sample values was induced by adding a common random component to each sample value, using $X_1^/ = X_1 + cU/\sqrt{1+c^2}$ and $X_2^/ = X_2 + cU/\sqrt{1+c^2}$, where U is a unit normal deviate, $c = \sqrt{\rho/(1-\rho)}$, and ρ is the desired correlation. If X_1 and X_2 are independent random variables with mean 0 and variance 1, then the correlation between $X_1^{'}$and $X_2^{'}$ is ρ. There were 50,000 iterations of the sampling procedure for each condition in Tables 2, 3, 4, and 5, where sample sizes were large, and 100,000 iterations for each condition in Tables 6 and 7, where sample sizes were smaller. There were 20,000 iterations for each point plotted in the figures. All significance tests were non-directional, except for the ones represented in Tables 3 and 5.

As a check, some of the simulations were repeated using the random number generator introduced by Marsaglia, Zaman, and Tsang (1990), described by Pashley (1993, pp. 395-415). Normal variates, N(0,1), were generated by the rejection method of Marsaglia and Bray (1964) and were transformed to have various distribution shapes using inverse distribution functions. The results of these methods were extremely close to the values in Tables 3 and 5, so all subsequent random deviates were obtained directly from *Mathematica* statistical add-on packages. For further details

[1] Copies of the *Mathematica* code used in this study can be obtained by writing to the author.

concerning simulation of non-normal variates, see, for example, Dagpunar (2007), Evans, Hastings & Peacock (2000), Gentle (1998), and Robert & Casella (2004).

SIMULATION RESULTS

The upper sections of Figure 1 compare the Type I error rates of the *z* test and the "*z* test for correlated samples," as a function of correlation, for sample sizes of 20 and 100. The lower sections compare the independent-samples *t* test and the modified *t* test using a correction for correlation for the same sample sizes. The Type I error rates of the conventional *z* test are seriously disrupted by correlation. For positive population correlations, the probability of rejecting H_0 falls below the .05 significance level and continues to decline as the correlation increases. For negative population correlations, the probability of rejecting H_0 exceeds the .05 level.

The discrepancies can be attributed to overestimation or underestimation of the standard error of the mean in the denominator of the *z* statistic when the correlation term in equation (1) is ignored. Using the modified estimate that includes the correlation term restores the probability to values very close to the .05 level, and the results are the same for both sample sizes.

Apparently the outcome is quite similar in the case of the independent-samples *t* test and the modified *t* test with a correlation term. Again, overestimation or underestimation of the standard error in the denominator of the *t* statistic that results from using equation (2) apparently accounts for the difference. In the case of $N = 20$, the probabilities based on the modified statistic do not quite reach the .05 significance level, although the correction improved for $N = 100$.

Table 2 compares the Type I error probabilities and power of three significance tests, for normal distributions, when population correlations ranged from $-.60$ to .60 in increments of .30. Sample sizes were 25, 100, and 400, and the significance levels were .05 and .01. In the second column, the difference between means was expressed as 0, 1, 2, and 3 times a fixed value based on the standard error of the mean. For N's of 25, 100, and 400, these values were .4, .2, and .1, respectively. Because of this adjustment, the probability values for each statistic turned out to be similar for all three sample sizes.

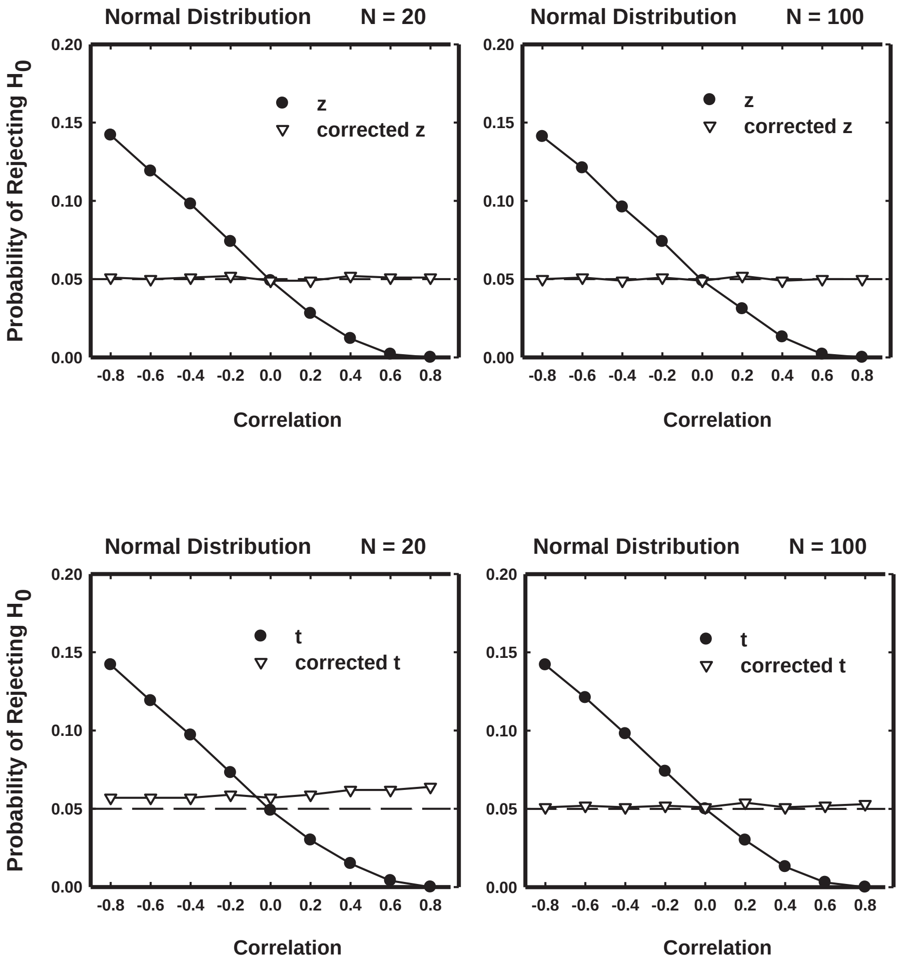


Figure 1. Probability of rejecting H_0 by z and t tests as a function of correlation for normal distribution (α = .05, N = 20 and 100).

Table 2. Type I error probabilities and power of the independent-samples Student t test (t), modified t test (tc), and paired-samples t test using difference scores (tp), nondirectional tests.

α = .05		*N* = 25			*N* = 100			*N* = 400		
ρ	*δ*	*t*	*tc*	*tp*	*t*	*tc*	*tp*	*t*	*tc*	*tp*
−.60	0	.120	.056	.047	.121	.053	.050	.120	.049	.049
	1	.314	.196	.176	.322	.195	.190	.330	.199	.197
	2	.735	.577	.543	.747	.596	.588	.752	.604	.602
	3	.966	.908	.891	.968	.920	.917	.968	.920	.919
−.30	0	.082	.055	.046	.086	.053	.050	.086	.049	.049
	1	.301	.233	.207	.308	.232	.227	.235	.235	.235
	2	.759	.672	.638	.774	.694	.686	.776	.697	.695
	3	.978	.956	.947	.979	.961	.959	.978	.962	.961
0	0	.052	.058	.047	.053	.055	.052	.051	.051	.050
	1	.282	.289	.260	.288	.291	.283	.293	.293	.292
	2	.793	.791	.762	.807	.807	.801	.810	.809	.808
	3	.988	.987	.982	.988	.988	.987	.988	.988	.988
.30	0	.023	.060	.049	.020	.052	.049	.020	.052	.051
	1	.246	.389	.352	.255	.392	.383	.258	.395	.393
	2	.832	.913	.894	.846	.920	.916	.851	.923	.922
	3	.996	.999	.999	.997	.999	.999	.997	.999	.999
.60	0	.004	.059	.047	.003	.052	.049	.002	.050	.049
	1	.194	.598	.555	.191	.606	.596	.035	.367	.363
	2	.894	.993	.990	.911	.993	.993	.915	.993	.993
	3	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
α = .01										
−.60	0	.042	.015	.011	.042	.013	.012	.041	.010	.010
	1	.169	.081	.164	.173	.077	.072	.178	.074	.072
	2	.535	.333	.281	.563	.355	.342	.572	.359	.356
	3	.892	.736	.682	.906	.770	.758	.907	.775	.772
−.30	0	.024	.015	.010	.025	.011	.010	.023	.010	.010
	1	.144	.097	.076	.151	.094	.088	.091	.091	.092
	2	.542	.422	.363	.578	.452	.438	.584	.459	.456
	3	.914	.840	.796	.926	.865	.857	.931	.874	.872
0	0	.011	.014	.009	.012	.012	.011	.011	.011	.011
	1	.113	.124	.095	.122	.125	.117	.120	.122	.120
	2	.561	.564	.499	.591	.591	.576	.599	.599	.595
	3	.940	.933	.907	.950	.948	.944	.952	.951	.951
.30	0	.003	.015	.010	.003	.012	.010	.002	.011	.011
	1	.079	.186	.147	.082	.189	.177	.082	.188	.185
	2	.570	.754	.694	.607	.781	.768	.617	.790	.787
	3	.965	.990	.983	.974	.993	.992	.977	.994	.994
.60	0	.000	.015	.010	.000	.012	.009	.000	.011	.010
	1	.040	.357	.295	.034	.362	.345	.035	.367	.363
	2	.592	.960	.939	.641	.969	.966	.654	.970	.970
	3	.986	1.000	1.000	.994	1.000	1.000	.996	1.000	1.000

For negative correlations, the Type I error rates and power of the independent-samples t test were both spuriously elevated, and for positive correlations, the probabilities declined below the .05 level. The discrepancies varied inversely with the degree of correlation and was close to zero for zero correlation. These results were independent of the sample size, for all degrees of correlation and for both significance levels. When $\rho < 0$, the power values are reduced below the values when $\rho = 0$, and when $\rho > 0$, the power values are inflated. The adequacy of the correction formulas is found by comparing of the corrected values with the corresponding entries the table for independent samples when $\rho = 0$.

As expected, the paired-samples *t* test performed well despite sizeable correlations. In all cases, the Type I error probabilities were restored to values close to the nominal significance level. The power increased in the case of positive correlations, and the spuriously large probabilities of rejecting H_0 were reduced in the case of negative correlations. Again, the result was the same for all degrees of correlation and both significance levels.

The pattern of results for the modified *t* test was quite similar to that for the paired-samples *t* test. The modified test evidently protected the Type I error probability very well in the case of sample sizes of 100 and 400 and to some degree for the sample size of 25. Furthermore, the modified *t* test showed a very slight but consistent increase in power compared to the paired-samples test, for $N = 100$. For $N = 400$, the power values were almost identical for the two tests. In the case of $N = 25$, the modified test showed an apparent increase in power, but, at the same time, the Type I error probabilities were somewhat elevated, so that the meaning of this result is questionable.

Table 3 provides similar data for directional tests, using the same values of N and ϱ in Table 2. Apparently the conclusions are the same, although the power values increased as expected, for both the .05 and .01 significance levels. The Type I error probabilities remained about the same and are close to the nominal significance levels.

These results are consistent with the advantage expected of a test based on $N_1 + N_2 - 2$ degrees of freedom compared to one based on $N - 1$ degrees of freedom. The fact that the difference decreases as sample size becomes large (e.g., $N = 400$) also is consistent with the same interpretation. The inaccuracy of the modified *t* test for $N = 25$ and its improvement as sample sizes become larger can be explained by the dependence of the variability of the sample correlation coefficients on sample size.

Table 3. Type I error probabilities and power of the independent-samples Student t test (t) , modified t test (tc), and paired-samples t test using difference scores (tp), directional tests.

α = .05		*N* = 25			*N* = 100			*N* = 400		
ρ	*δ*	*t*	*tc*	*tp*	*t*	*tc*	*tp*	*t*	*tc*	*tp*
−.60	0	.102	.061	.055	.103	.057	.056	.096	.051	.051
	1	.404	.287	.266	.416	.290	.284	.420	.298	.297
	2	.817	.702	.677	.824	.717	.711	.823	.718	.717
	3	.983	.955	.947	.985	.960	.958	.982	.958	.958
−.30	0	.077	.058	.051	.078	.054	.053	.074	.051	.051
	1	.402	.335	.312	.411	.337	.331	.415	.339	.338
	2	.842	.785	.764	.848	.793	.788	.856	.803	.802
	3	.991	.982	.977	.990	.982	.981	.990	.983	.983
0	0	.053	.056	.050	.049	.051	.049	.049	.049	.049
	1	.394	.399	.373	.403	.404	.398	.409	.409	.408
	2	.876	.875	.860	.881	.880	.877	.883	.884	.883
	3	.994	.994	.993	.994	.994	.994	.995	.995	.995
.30	0	.028	.056	.049	.027	.052	.051	.024	.048	.048
	1	.376	.508	.480	.394	.520	.521	.392	.516	.515
	2	.914	.953	.946	.918	.955	.953	.918	.957	.957
	3	.999	1.000	1.000	.999	.999	.999	.998	.999	.999
.60	0	.007	.050	.044	.005	.053	.051	.005.	.050	.049
	1	.351	.715	.690	.353	.719	.713	354	.716	.714
	2	.965	.997	.997	.964	.997	.997	.967	.997	.997
	3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
α = .01										
−.60	0	.040	.018	.015	.038	.015	.014	.036	.012	.012
	1	.215	.117	.096	.225	.115	.111	.236	.114	.113
	2	.618	.432	.382	.648	.453	.442	.651	.457	.455
	3	.929	.820	.784	.936	.840	.833	.935	.845	.844
−.30	0	.025	.016	.013	.024	.014	.013	.022	.011	.011
	1	.200	.142	.117	.203	.136	.130	.209	.140	.138
	2	.642	.530	.478	.655	.548	.534	.670	.560	.557
	3	.949	.899	.873	.954	.918	.912	.956	.920	.919
0	0	.012	.014	.011	.011	.012	.011	.011	.011	.011
	1	.167.	.176	.148	.177	.178	.170	.180	.180	.178
	2	.661	.661	.615	.687	.686	.674	.694	.693	.690
	3	.963	..960	.947	.970	.969	.967	.973	.973	.972
.30	0	.004	.014	.011	.003	.012	.011	.004	.011	.011
	1	.129	.252	.217	.138	.264	.255	.141	.263	.260
	2	.691	.830	.795	.715	.846	.839	.727	.850	.848
	3	.984	.994	.992	.986	.996	.995	.987	.996	.996
.60	0	.000	.012	.010	.000	.011	.010	.005	.050	.049
	1	.081	.446	.396	.077	.454	.441	.354	.716	.714
	2	.741	.976	.969	.771	.981	.980	.967	.997	.997
	3	.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4 presents similar comparisons of three nonparametric tests applied to paired-samples data from a normal and nine non-normal distributions. All sample sizes were 100, and the population correlation was .40. Many previous simulation studies have shown that the Wilcoxon-Mann-Whitney test is substantially superior to the independent-samples *t* test from several non-normal distributions in the table. Studies have also shown that the Wilcoxon signed-ranks test is superior to the paired-samples *t* test for the same distributions. Table 5 presents similar results for directional tests.

It is clear that the Type I error probabilities of the Wilcoxon-Mann-Whitney test fall considerably below the nominal significance levels as a result of the correlation. However, the Wilcoxon signed-ranks test restores the significance level and substantially increases power, just as the paired-sample *t* test does in the corresponding case of samples from a normal distribution.

For these non-normal distributions, the modified *t* test on rank-transformed data performed just as well as the Wilcoxon signed-ranks test and in some cases was superior to the Wilcoxon test, especially for the .01 significance level. It appears, therefore, that the modified *t* test on ranks combines the advantage of a modified correction for correlation for paired data, using $N_1 + N_2 - 2$ instead of $N - 1$ degrees of freedom, with the advantage of a nonparametric test for non-normal distributions. Tables 6 and 7 provide similar information for smaller sample sizes that are widely employed in research studies. In most cases, the differences between the three significance tests is larger for these small *N*'s. Also, the elevation of the Type I error probability of the modified test above the nominal significance level is somewhat larger. Again, this result is consistent with the increased variability of the sample correlation coefficient for small sample sizes.

Figure 2 provides more detailed power functions of the three significance tests for normal distributions. When $\rho = 0$ (upper section), the functions for the independent-samples t and corrected *t* are nearly identical, while the paired-samples t was somewhat less powerful for all differences between means. Again, this outcome is consistent with the slightly increased power that would be expected from the difference between $N_1 + N_2 - 2$ degrees of freedom and $N - 1$ degrees of freedom. When $\rho =$ .40 (lower section), both the paired-samples *t* and the modified *t* were considerably more powerful than the independent-samples *t* test as is expected. Moreover, the corrected *t* is slightly more powerful than the paired-samples *t*, again consistent with the difference in degrees of freedom.

Table 4. Type I error probabilities and power of Wilcoxon-Mann-Whitney test (W), modified t test on ranks (tc_R), and Wilcoxon signed-ranks test (WS), N = 100, ρ = .40, nondirectional tests.

		α = .05			α = .01		
Distribution	*δ*	*W*	*$(tc)_R$*	*WS*	*W*	*$(tc)_R$*	*WS*
	0	.022	.051	.049	.003	.012	.011
normal	1	.242	.367	.367	.077	.171	.165
	2	.827	.903	.906	.578	.749	.744
	3	.995	.998	.998	.964	.988	.989
	0	.018	.050	.049	.002	.010	.009
exponential	1	.402	.571	.524	.162	.334	.282
	2	.962	.984	.973	.850	.940	.901
	3	1.000	.1.000	1.000	.998	1.000	.998
	0	.017	.051	.047	.002	.011	.009
Laplace	1	.306	.463	.439	.109	.235	.213
	2	.905	.954	.944	.719	.861	.833
	3	.999	1.000	.999	.990	.997	.995
	0	.017	.052	.048	.002	.011	.010
lognormal	1	.773	.881	.824	.496	.719	.615
	2	.999	1.000	.999	.994	.998	.993
	3	1.000	1.000	1.000	1.000	1.000	1.000
	0	.020	.052	.049	.003	.011	.010
logistic	1	.267	.403	.397	.089	.196	.185
	2	.858	.924	.922	.631	.793	.779
	3	.997	.999	.999	.977	.993	.992
	0	.019	.051	.049	.003	.011	.010
half-normal	1	.266	.407	.390	.085	.197	.179
	2	.859	.926	.919	.635	.799	.774
	3	.997	.999	.999	.979	.993	.991
	0	.021	.052	.049	.002	.011	.009
uniform	1	.226	.340	.351	.069	.153	.153
	2	.783	.869	.883	.526	.694	.706
	3	.990	.996	.997	.945	.981	.983
mixed-normal	0	.019	.051	.047	.002	.011	.009
p = .02	1	.332	.477	.473	.124	.252	.239
k = 5	2	.927	.966	.965	.764	.885	.876
	3	1.000	1.000	1.000	.995	.999	.999
	0	.017	.050	.048	.002	.010	.009
chi-square	1	.340	.494	.469	.123	.267	.239
	2	.928	.968	.957	.768	.894	.856
	3	1.000	1.000	1.000	.994	.999	.997
	0	.020	.052	.049	.002	.011	.010
extreme value	1	.285	.429	.416	.097	.211	.198
(Gumbel)	2	.880	.941	.934	.672	.828	.806
	3	.998	.999	.999	.985	.996	.995

Table 5. Type I error probabilities and power of Wilcoxon-Mann-Whitney test (W), modified t test on ranks (tc_R), and Wilcoxon signed-ranks test (WS), N = 100, ρ = .40, directional tests.

		α = .05			α = .01		
Distribution	*δ*	*W*	*$(tc)_R$*	*WS*	*W*	*$(tc)_R$*	*WS*
normal	0	.019	.054	.052	.002	.012	.010
	1	.364	.551	.554	.112	.291	.280
	2	.919	.968	.970	.698	.880	.882
	3	.999	1.000	1.000	.989	.998	.998
exponential	0	.017	.051	.052	.002	.011	.010
	1	.545	.718	.702	.223	.470	.437
	2	.986	.995	.993	.909	.975	.965
	3	1.000	1.000	1.000	.999	1.000	1.000
Laplace	0	.018	.053	.054	.001	.012	.010
	1	.442	.629	.626	.154	.367	.350
	2	.961	.987	.986	.820	.940	.936
	3	1.000	1.000	1.000	.998	1.000	1.000
lognormal	0	.018	.056	.048	.001	.015	.011
	1	.869	.939	.923	.602	.818	.778
	2	1.000	1.000	1.000	.998	.999	.999
	3	1.000	1.000	1.000	1.000	1.000	1.000
logistic	0	.018	.051	.048	.002	.011	.009
	1	.395	.582	.581	.125	.321	.304
	2	.938	.978	.977	.751	.907	.904
	3	1.000	1.000	1.000	.992	.999	.999
half-normal	0	.018	.052	.048	.001	.011	.010
	1	.392	.575	.574	.128	.320	.307
	2	.939	.979	.978	.751	.908	.903
	3	.999	1.000	1.000	.992	.998	.999
uniform	0	.020	.056	.052	.002	.011	.010
	1	.340	.529	.526	.099	.272	.262
	2	.901	.961	.962	.660	.860	.859
	3	.999	1.000	1.000	.981	.997	.998
mixed-normal	0	.019	.057	.052	.001	.014	.011
p = .02	1	.474	.658	.656	.177	.400	.382
k = 5	2	.974	.991	.992	.860	.958	.958
	3	1.000	1.000	1.000	.998	1.000	1.000
chi-square	0	.018	.055	.052	.002	.012	.010
	1	.475	.660	.648	.174	.398	.374
	2	.974	.992	.991	.854	.956	.950
	3	1.000	1.000	1.000	.999	1.000	1.000
extreme value (Gumbel)	0	.018	.053	.051	.001	.012	.010
	1	.416	.605	.599	.136	.344	.326
	2	.954	.984	.984	.789	.928	.926
	3	1.000	1.000	1.000	.995	.999	.999

Table 6. Type I error probabilities and power of independent-samples Student t test (t), modified t test (tc), and paired-samples t test using difference scores (tp), for small samples from normal distribution.

α = .05		*N* = 8			*N* = 10			*N* = 15		
ρ	*δ*	*t*	*tc*	*tp*	*t*	*tc*	*tp*	*t*	*tc*	*tp*
0	0	.051	.071	.040	.049	.066	.040	.050	.061	.044
	1	.153	.185	.120	.182	.182	.127	.256	.271	.226
	2	.453	.476	.358	.548	.565	.458	.749	.747	.690
	3	.794	.786	.675	.891	.900	.838	.980	.976	.964
.30	0	.023	.076	.039	.023	.069	.040	.022	.064	.044
	1	.125	.238	.155	.142	.250	.173	.224	.362	.302
	2	.447	.613	.479	.560	.730	.621	.785	.882	.842
	3	.830	.907	.826	.933	.976	.948	.992	.997	.995
.60	0	.006	.080	.039	.005	.072	.040	.004	.066	.044
	1	.085	.362	.242	.096	.404	.291	.178	.563	.488
	2	.449	.841	.723	.588	.937	.878	.843	.987	.978
	3	.877	.991	.970	.966	.999	.998	.999	1.000	1.000
α = .01										
0	0	.011	.024	.008	.011	.019	.007	.011	.016	.009
	1	.050	.078	.033	.067	.068	.032	.100	.119	.077
	2	.210	.255	.130	.285	.299	.168	.493	.504	.398
	3	.516	.540	.334	.671	.690	.502	.902	.891	.822
.30	0	.004	.026	.008	.004	.021	.008	.003	.018	.009
	1	.034	.109	.046	.039	.102	.048	.073	.175	.116
	2	.184	.368	.194	.251	.457	.278	.491	.689	.578
	3	.527	.719	.495	.705	.869	.718	.933	.976	.948
.60	0	.001	.028	.008	.000	.023	.007	.000	.019	.009
	1	.017	.185	.078	.018	.191	.089	.040	.326	.227
	2	.156	.621	.377	.230	.768	.568	.500	.930	.870
	3	.544	.933	.787	.745	.990	.956	.965	1.000	.999

Table 7. Type I error probabilities and power of Wilcoxon-Mann-Whitney test (W), modified t test on ranks (tc_R) , and Wilcoxon signed-ranks test (WS), for non-normal distributions (N = 10, ρ = .40).

		α = .05			α = .01		
Distribution	δ	*W*	*(tc)$_R$*	*WS*	*W*	*(tc)$_R$*	*WS*
normal	0	.019	.068	.048	.003	.020	.007
	1	.130	.263	.214	.034	.109	.038
	2	.558	.764	.720	.252	.506	.261
	3	.934	.982	.975	.712	.902	.690
exponential	0	.016	.066	.049	.002	.019	.006
	1	.214	.416	.326	.062	.225	.104
	2	.703	.841	.743	.417	.667	.414
	3	.937	.972	.932	.783	.907	.718
Laplace	0	.017	.068	.049	.002	.019	.005
	1	.172	.344	.283	.049	.172	.077
	2	.637	.801	.734	.345	.598	.362
	3	.931	.971	.945	.752	.897	.704
lognormal	0	.012	.062	.048	.002	.017	.005
	1	.424	.640	.470	.177	.416	.202
	2	.875	.947	.798	.654	.843	.582
	3	.977	.991	.912	.890	.958	.799
logistic	0	.018	.070	.049	.003	.021	.006
	1	.150	.307	.260	.039	.145	.060
	2	.593	.769	.722	.297	.552	.322
	3	.924	.971	.955	.726	.888	.699
half-normal	0	.016	.062	.043	.002	.017	.004
	1	.145	.301	.244	.037	.136	.053
	2	.588	.773	.715	.287	.546	.307
	3	.920	.969	.956	.716	.887	.676
uniform	0	.018	.071	.048	.003	.022	.006
	1	.128	.267	.229	.031	.119	.047
	2	.531	.715	.694	.239	.470	.258
	3	.907	.971	.966	.683	.872	.666
mixed-normal p = .02 k = 5	0	.018	.069	.050	.002	.021	.006
	1	.184	.355	.304	.052	.175	.075
	2	.703	.849	.793	.396	.654	.420
	3	.965	.986	.942	.831	.941	.791
chi-square	0	.016	.066	.048	.002	.019	.006
	1	.184	.372	.304	.051	.191	.086
	2	.664	.821	.737	.362	.629	.383
	3	.940	.976	.943	.772	.907	.710
extreme value (Gumbel)	0	.017	.069	.049	.002	.021	.006
	1	.159	.330	.272	.042	.159	.067
	2	.624	.790	.728	.321	.579	.343
	3	.929	.972	.950	.749	.897	.704

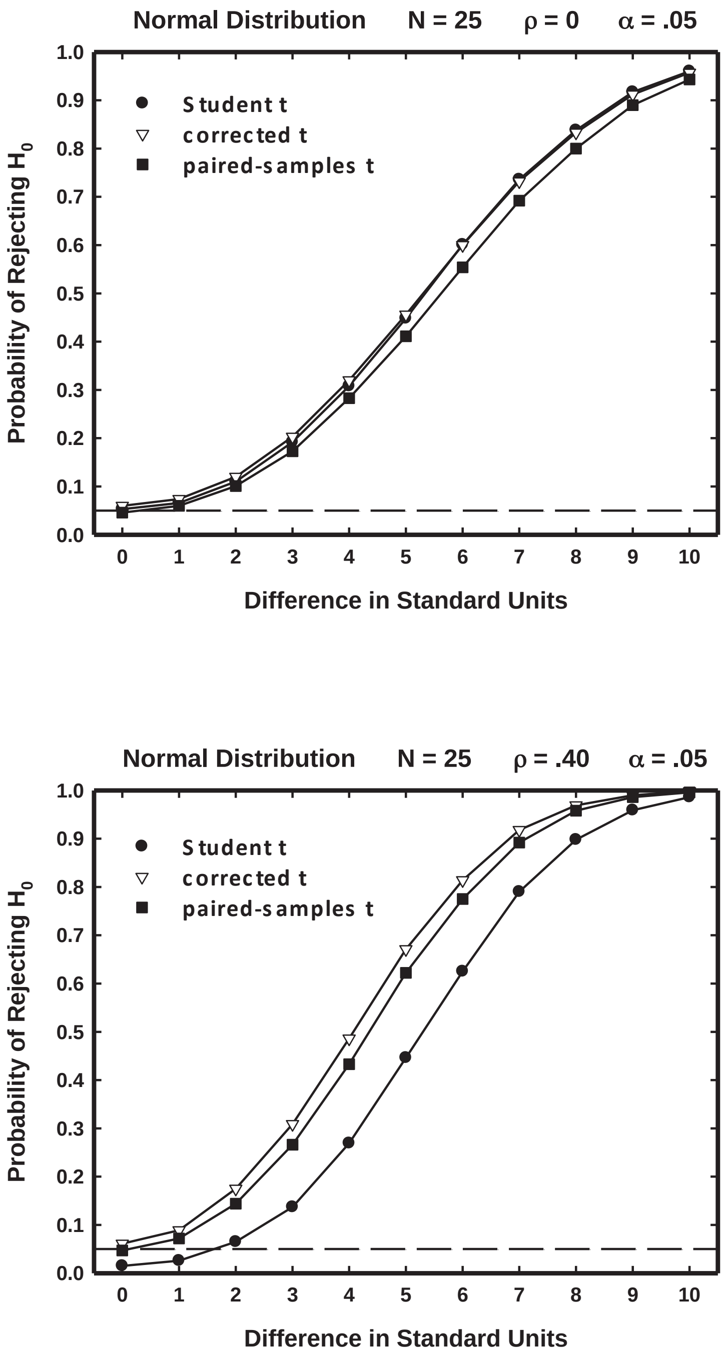


Figure 2. Power functions for normal distribution (N = 25, α = .05, $\rho = 0$ and .40).

Figure 3 shows similar power functions for exponential distributions, where the Wilcoxon rank-sum test, the Wilcoxon signed-ranks test, and the modified *t* test on ranks were substituted for the parametric *t* tests. Comparison of the three power functions reveals that the outcome is almost the same as in the case of the corresponding parametric tests applied to normally distributed data. The Wilcoxon signed-ranks test was superior to the Wilcoxon rank-sum test for paired data, while the modified *t* test on ranks was slightly superior to both.

Apparently the modified *t* test corrected for the correlation resulting from pairing, while at the same time the transformation to ranks counteracted non-normality. Figures 4, 5, and 6 indicate similar outcomes for lognormal, chi-square, half-normal, and uniform distributions, using several sample sizes, population correlations, and significance levels. Note that the power functions for the smaller sample sizes were more widely separated, while convergence is evident for the larger sample sizes.

Table 8 compares Type I error probabilities when a sample correlation is entered into equation (2) for each sample taken and when a fixed population correlation is entered the equation for every sample. The first section of the table, for the normal distribution, is the result of the *t* test performed on scores. The remaining three sections, for non-normal distributions, show the result of the *t* test on rank-transformed data. For relatively small correlations and relatively large sample sizes, the Type I error probabilities for both tests were about the same and close to the nominal significance level.

SOME PRACTICAL IMPLICATIONS

For samples of size 25 or 50 from normal distributions, the modified *t* test with a correction for correlation maintained Type I error rates close to the significance level, increased power in the case of positive correlations, and removed spurious increases in the probability of rejecting H_0 in the case of negative correlations. The power superiority of this test over the paired-samples *t* test is about what one would expect from the difference in degrees of freedom. The difference became less marked as sample sizes increased to 100 and 400, presumably because the difference in the critical values of the *t* statistic for $N - 1$ and $2N - 2$ degrees of freedom decreases as N increases. Nevertheless, the power of the modified test was equal to that of the paired-samples test for the larger sample sizes.

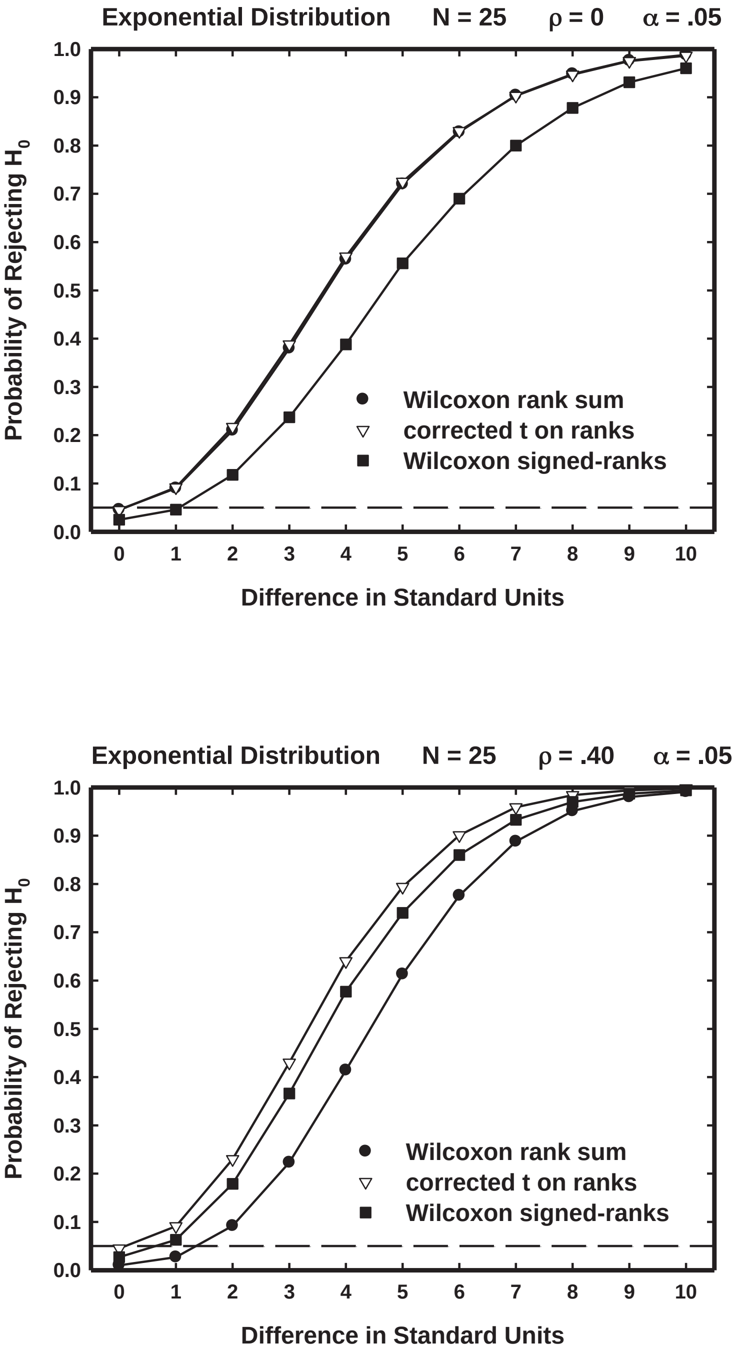


Figure 3. Power functions for exponential distribution (N = 25, α = .05, $\rho = 0$ and .40).

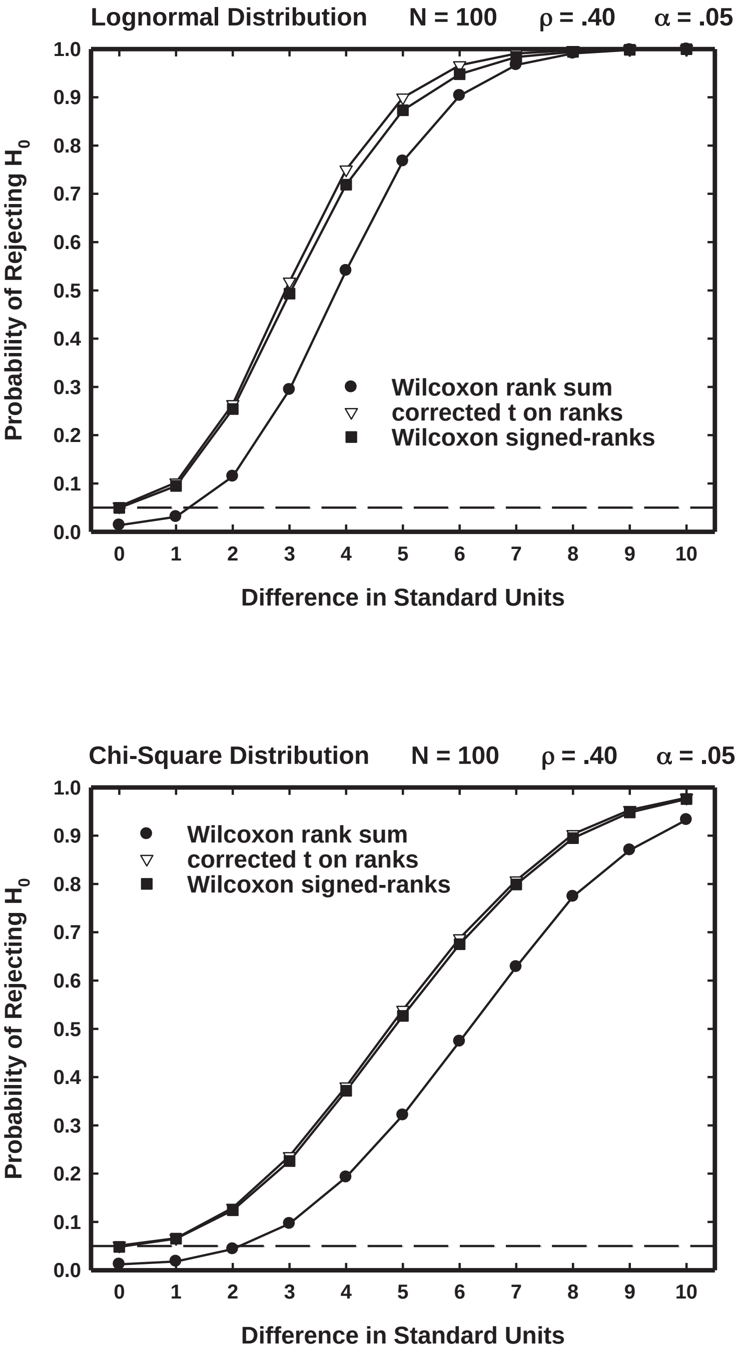


Figure 4. Power functions for lognormal and chi-square distributions (N = 100, α = .05, ρ =.40).

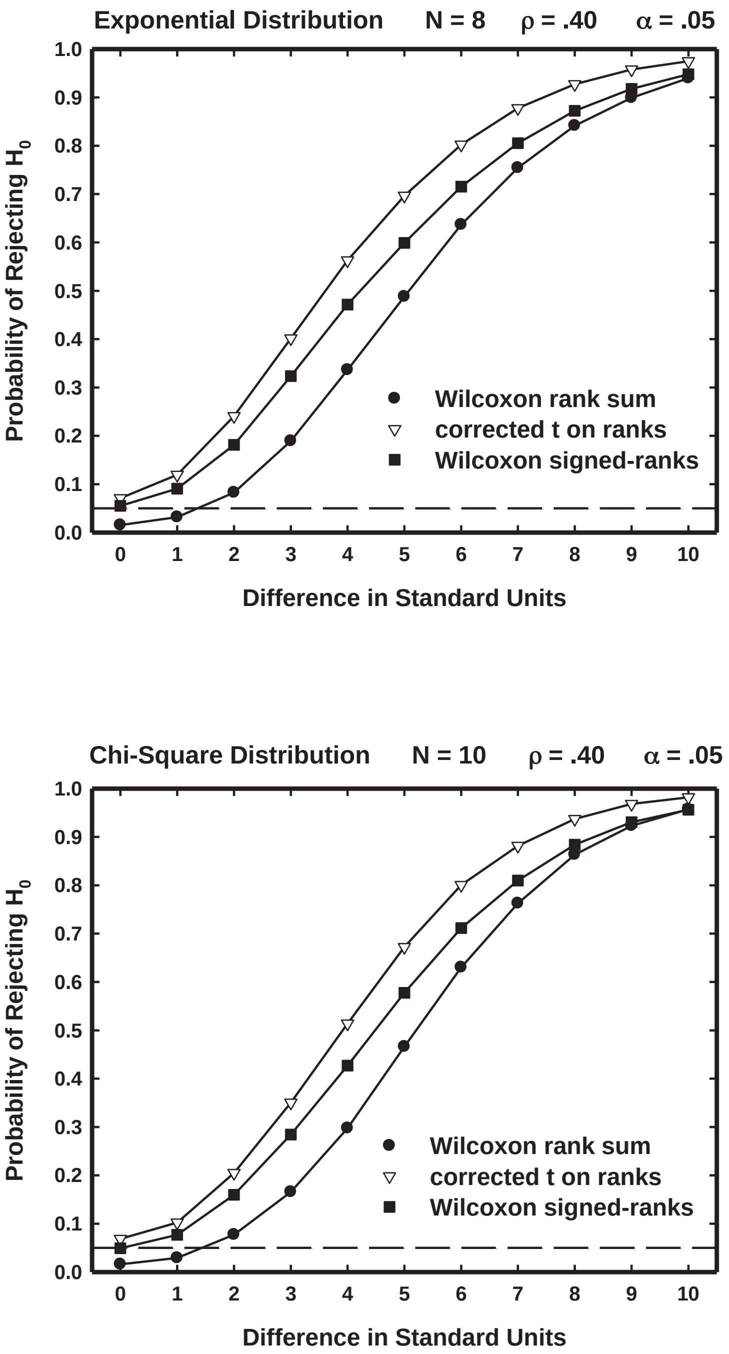
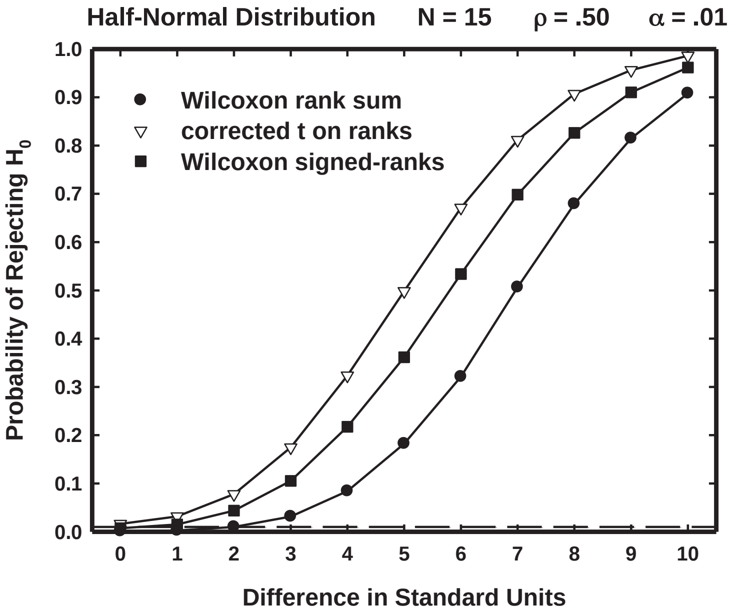


Figure 5. Power functions for exponential distribution (N = 8, α = .05, ρ= .40) and chi-square distribution (N = 10, α = .05, ρ = .40).



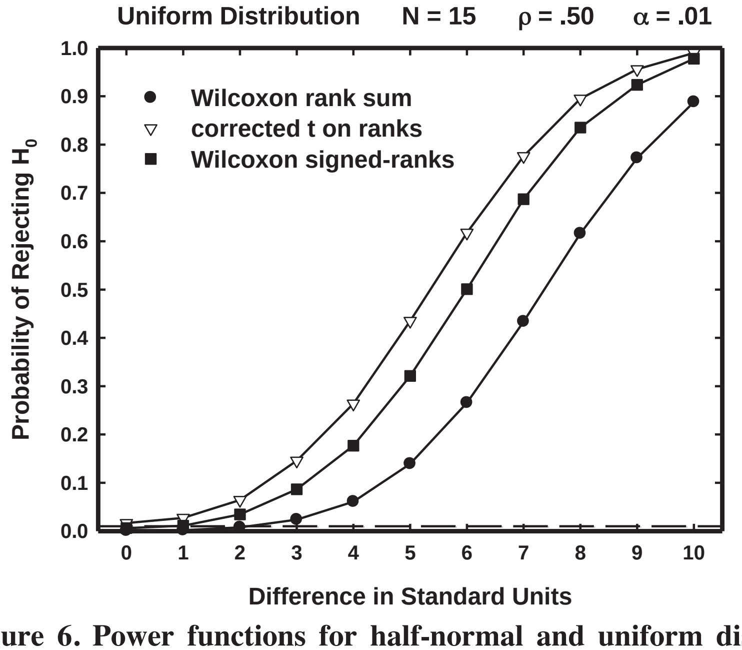


Figure 6. Power functions for half-normal and uniform distributions (N = 15, α = .01, ρ = .50).

Table 8. Type I error probabilities for corrections using sample correlations, c(r), and population correlations, c(ρ).

α = .05		*N* = 10		*N* = 25		*N* = 50		*N* = 100	
Distribution	*ρ*	*c(r)*	*c(ρ)*	*c(r)*	*c(ρ)*	*c(r)*	*c(ρ)*	*c(r)*	*c(ρ)*
	0	.064	.050	.056	.051	.055	.051	.052	.050
	.20	.065	.051	.057	.051	.053	.051	.052	.051
normal	.40	.065	.053	.057	.052	.055	.052	.052	.051
	.60	.067	.056	.056	.053	.053	.052	.051	.050
	.80	.068	.060	.057	.055	.054	.051	.053	.052
	0	.068	.052	.052	.045	.052	.049	.051	.050
	.20	.067	.044	.055	.041	.052	.043	.052	.045
exponential	.40	.065	.052	.054	.049	.052	.049	.050	.049
	.60	.067	.064	.055	.066	.052	.066	.051	.067
	.80	.068	.109	.056	.109	.054	.107	.051	.104
	0	.067	.052	.053	.046	.051	.049	.052	.051
	.20	.070	.049	.055	.045	.053	.047	.050	.046
Laplace	.40	.069	.057	.057	.051	.051	.050	.052	.052
	.60	.068	.061	.055	.063	.053	.065	.050	.062
	.80	.069	.096	.056	.093	.052	.089	.051	.087
	0	.067	.053	.051	.044	.052	.049	.050	.049
	.20	.071	.055	.055	.050	.051	.052	.050	.052
uniform	.40	.072	.060	.055	.054	.052	.054	.052	.054
	.60	.074	.053	.058	.053	.053	.052	.051	.050
	.80	.074	.063	.059	.057	.052	.048	.051	.046
α = .01									
	0	.018	.011	.014	.011	.013	.012	.012	.011
	.20	.019	.011	.014	.011	.013	.011	.012	.011
normal	.40	.018	.012	.015	.012	.013	.012	.012	.011
	.60	.020	.013	.014	.012	.012	.012	.012	.011
	.80	.021	.016	.014	.012	.013	.012	.012	.012
	0	.021	.011	.012	.010	.011	.010	.011	.010
	.20	.018	.010	.013	.008	.011	.008	.011	.008
exponential	.40	.018	.012	.012	.009	.011	.010	.010	.010
	.60	.018	.018	.012	.015	.010	.016	.011	.016
	.80	.018	.039	.012	.033	.011	.033	.011	.033
	0	.021	.012	.013	.010	.011	.009	.011	.010
	.20	.021	.012	.013	.009	.012	.009	.010	.009
Laplace	.40	.020	.013	.013	.010	.011	.010	.011	.011
	.60	.019	.018	.012	.014	.011	.015	.010	.014
	.80	.019	.031	.012	.026	.011	.026	.010	.025
	0	.021	.011	.011	.018	.011	.010	.011	.010
	.20	.022	.015	.013	.011	.011	.011	.010	.011
uniform	.40	.023	.014	.014	.011	.011	.011	.011	.012
	.60	.024	.013	.014	.011	.011	.010	.010	.010
	.80	.024	.015	.015	.011	.011	.009	.011	.009

For small N's of 8, 10, and 15, the same differences in power functions were evident, but the interpretation of these differences is problematic, because the Type I error rates of the modified *t* test were somewhat higher than the nominal significance levels. Generally the Type I error rate was about .060 for the .05 significance level and about .014 for the .01 significance level. Possibly these disparities resulted from variability of the sample correlation coefficient for small *N*.

The elevation, rather than a depression of the probability of rejecting H_0 can be explained by the left-skewness of the distribution of the sample correlation coefficient for positive values of the population correlation. For those positive values of ϱ, proportionately more high values of the sample *r* appeared in the denominator of equation (2), resulting in an inflated *t* statistic. However, as sample size increased, the distribution of the sample *r* became more nearly symmetrical, and the inflation was not as large.

The skewness is evident in Figure 7, which shows distributions of the sample correlation coefficient under the conditions represented in Figure 1 and Table 2, when the sample sizes were 25 and 100 and the population correlation was .50 and .75. The sample correlations were substantially left-skewed for the smaller sample size and became more symmetrical and less variable when the sample size increased. For N = 25, there was considerable overlap of the two distributions of sample values for population correlations of .50 and .75, and for N = 100, the distributions were more widely separated.

Figure 8 plots relative frequency distributions of the values of the *t* statistic. All four graphs are for a normal distribution with N = 25. The first distribution, at the top, shows the independent-samples *t* statistics when $\rho = 0$. The second distribution shows a decrease in the variance of that distribution when $\rho = .50$. The remaining distributions are for the two methods of correcting for correlation based on *r* and ρ. The two distributions of the corrected statistics have nearly the same variance, and both restore the distributions close to their shape of the one in the graph at the top. Means and values of the distributions of the t statistics, the two corrections, and the paired-samples t statistic are shown in Table 9 for various sample sizes and population correlations.

For non-normal distributions, the results were similar. The Wilcoxon signed-ranks test is related to the Wilcoxon-Mann-Whitney rank sum test in the same way as the paired-samples *t* test is related to the independent-samples *t* test. However, there is no version of the Wilcoxon-Mann-Whitney test involving correlation coefficients corresponding to the *z* test for correlated samples. Since the Student *t* test with a rank transformation and the Wilcoxon-Mann-Whitney test are equivalent, the modified *t* test on ranks appears suitable in the case of paired data. This test preserved Type I error rates and increased power for the larger sample sizes. Again, there was an elevation of the probabilities of rejecting H_0 above the nominal significance level for the smaller sample sizes.

The modified *t* test on ranks performed about the same as both the paired-samples *t* test and the Wilcoxon signed-ranks test for small and moderate sample sizes, when the population correlation was used in the correction formula. However, in the case of small sample sizes, Type I error rates of the modified test were altered when sample correlations were used. For large sample sizes – 100 or more – all the tests performed about the same.

One might question, therefore, whether the advantage of acquiring more degrees of freedom is enough to outweigh the disadvantage of inflation of the Type I error rate for small sample sizes. Perhaps in some special circumstances the modified test could be advantageous. First, under some conditions, the population correlation coefficient between two paired groups may be known in advance. In a before-after experimental design, theory or previous research may have established the correlation between the pairs. In that case, the known value of ρ can be substituted into equation (2), and the variability of the sample *r* would be obviated, as suggested by the results in Table 8. For small sample sizes, the increase in power could be substantial. Although these special circumstances are unlikely in practical research, the modified *t* test can be a useful alternative to have available. Second, in the case of some non-normal data, an assumption of the Wilcoxon signed-ranks test, symmetry of the difference scores, may not be satisfied. In that case it is reasonable to employ the modified *t* test on ranks, which appears to be effective.

More recently, many additional statistical tests have been developed that are more accurate and more powerful than the traditional parametric and nonparametric methods listed in Table 1 (see, for example, Huber, 1996; Wilcox, 2003). The estimation of correlation has also improved in recent years (see, for example, Rousseeuw & Leroy, 1987; Wilcox & Muska, 2002; Zimmerman, Zumbo, & Williams, 2003). The modified *t* test of the present study is not a substitute for the best current statistical tests available, but is provided because of its theoretical interest and because it fills gaps in the classification of two-sample tests of location. Under conditions where limited computing resources are available, the correction for correlation could be useful as a practical method.

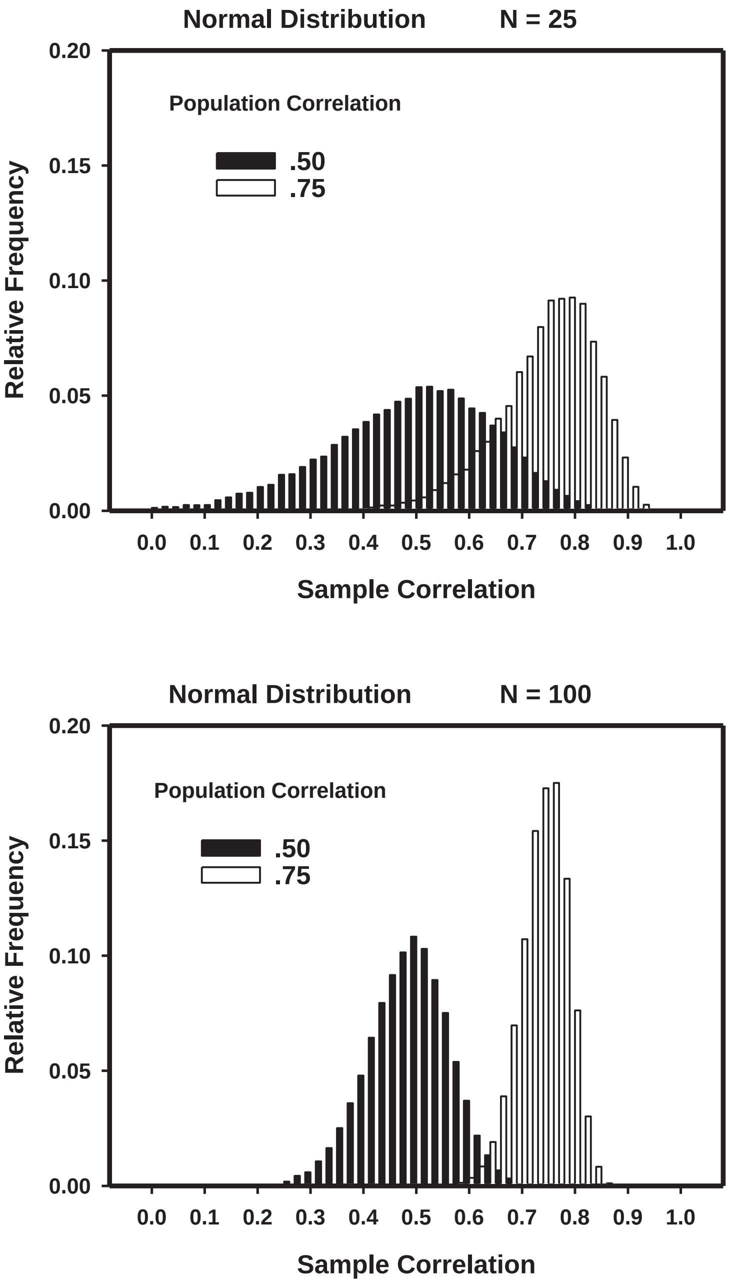


Figure 7. Relative frequency distributions of sample correlation coefficients for different values of the population correlation and sample size.

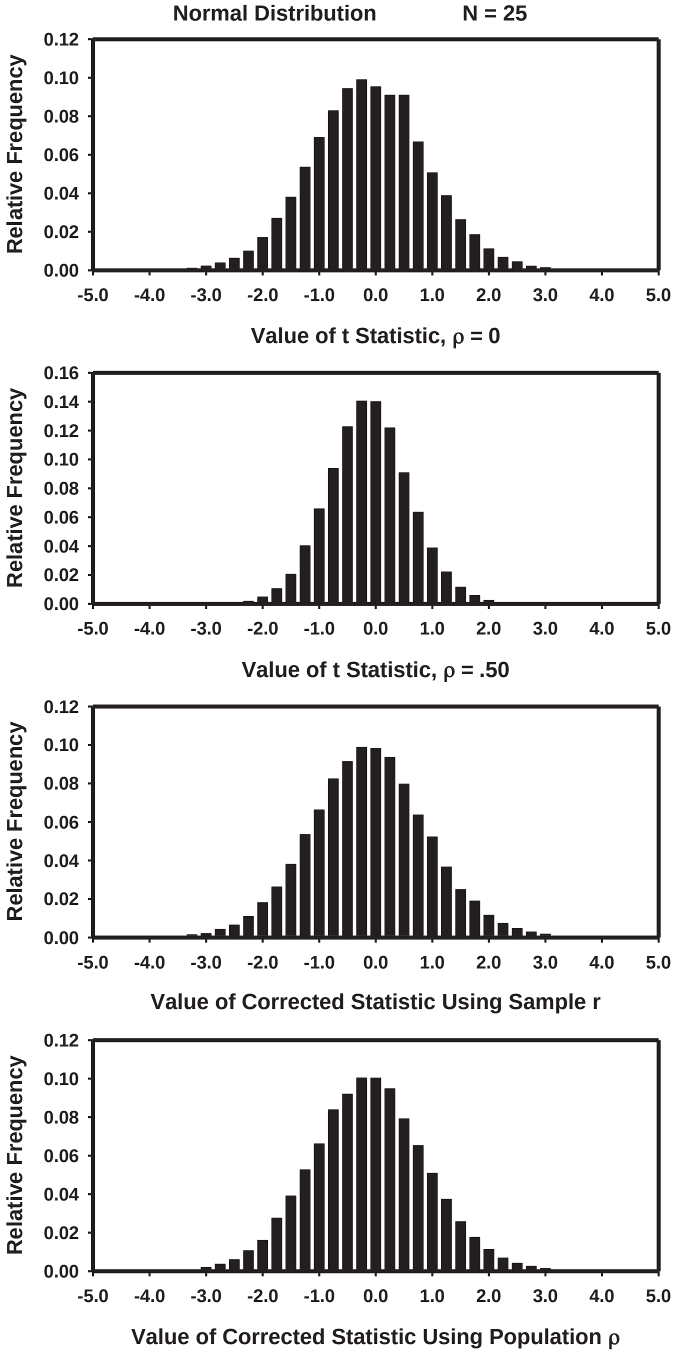


Figure 8. Relative frequency distributions of t statistics. corrected t statistics, and paired-samples t statistics.

Table 9. Means and standard deviations of uncorrected and corrected statistics. t_0: Student t statistic from uncorrelated pairs. t: Student t statistic from correlated pairs. t′ (using r): Corrected statistic using sample r in equation (2). t′ (using ρ): Corrected statistic using population ρ in equation (2).

		Population Correlation (ρ)					
N	Statistic	.25		.50		.75	
10	t_0	.001	1.060	.001	1.062	.004	1.067
	t	.002	.916	.001	.762	.000	.546
	t' (using r)	−.050	1.148	−.045	1.179	−.045	1.192
	t' (using ρ)	.002	1.078	.002	1.078	−.000	1.093
25	t_0	.003	1.022	−.003	1.018	−.001	1.031
	t	−.004	.890	.002	.726	.001	.519
	t' (using r)	−.003	1.052	.003	1.056	.005	1.066
	t' (using ρ)	−.005	1.027	.003	1.026	.003	1.037
50	t_0	.008	1.012	.003	1.007	−.007	1.008
	t	.002	.878	−.000	.717	.001	.507
	t' (using r)	.003	1.026	−.001	1.027	.003	1.028
	t' (using ρ)	.002	1.013	−.001	1.014	.002	1.014
100	t_0	.002	.999	−.001	1.002	−.001	1.004
	t	−.002	.869	.002	.711	−.004	.501
	t' (using r)	−.002	1.008	.002	1.012	−.009	1.008
	t' (using ρ)	−.002	1.003	.003	1.006	−.008	1.002

REFERENCES

Dagpunar, J.S. (2007). *Simulation and Monte Carlo*. New York: Wiley.

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions* (3rd ed.). New York: Wiley.

Gentle, J.E. (1998). *Random number generation and Monte Carlo methods*. New York: Springer.

Guilford, J.P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education* (5th ed.). pp. 154-155. New York: McGraw-Hill.

Hays, W.L. (1988). *Statistics* (4th ed). New York: Holt, Rinehart, & Winston.

Huber, P. (1996). *Robust statistical procedures* (2nd ed.). New York: Society for Industrial and Applied Mathematics.

Marsaglia, G., & Bray, T.A. (1964). A convenient method for generating normal variables. *Society for Industrial and Applied Mathematics Review, 6,* 260-264.

Marsaglia, G., Zaman, A., & Tsang, W.W. (1990). Toward a universal random number generator. *Statistics & Probability Letters, 8,* 35-39.

McNemar, Q. (1955). *Psychological statistics (*2nd ed.). New York: Wiley.

Pashley, P.J. (1993). On generating random sequences. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 395-415). Hillsdale, NJ: Lawrence Erlbaum Associates.

Robert, C.P., & Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.).New York: Springer-Verlag.

Rousseeuw, P.J., & Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Snedecor, G.W., & Cochran, G.W. (1989). *Statistical methods* (8th ed.). Ames, IA: Iowa State University Press.

Wilcox, R.R. (2003). *Applying contemporary statistical techniques*. New York: Academic Press.

Wilcox, R.R., & Muska (2002). Comparing correlation coefficients. *Communications in Statistics—Simulation and Computation, 31,* 49-59.

Zimmerman, D.W. (1997). A note on the paired-samples *t* test. *Journal of Educational and Behavioral Statistics, 22,* 349-360.

Zimmerman, D.W., Williams, R.H., & Zumbo, B.D. (1993). Effect of nonindependence of sample observations on parametric and nonparametric statistical tests. *Communications in Statistics: Simulation and Computation, 22,* 779-789.

Zimmerman, D.W., Zumbo, B.D., & Williams, R.H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicologica, 24,* 133-158.

(Manuscript received: 2 November 2011; accepted: 19 January 2012)