Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 22

11-1-2003

Conventional And Robust Paired And Independent-Samples *t* Tests: Type I Error And Power Rates

Katherine Fradette
University of Manitoba, umfradet@cc.umanitoba.ca

H. J. Keselman
University of Manitoba, kesel@ms.umanitoba.ca

Lisa Lix
University of Manitoba, lisa.lix@usask.ca

James Algina
University of Florida, algina@ufl.edu

Rand R. Wilcox
University of Southern California, rwilcox@usc.edu

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 481-496

1538 – 9472/03/$30.00

Early Scholars

Conventional And Robust Paired And Independent-Samples *t* Tests: Type I Error And Power Rates

Katherine Fradette and H. J. Keselman
University of Manitoba
Department of Psychology

Lisa Lix
University of Manitoba
Department of Community Health Sciences

James Algina
University of Florida
Department of Educational Psychology

Rand R. Wilcox
University of Southern California
Department of Psychology

Monte Carlo methods were used to examine Type I error and power rates of 2 versions (conventional and robust) of the paired and independent-samples t tests under nonnormality. The conventional (robust) versions employed least squares means and variances (trimmed means and Winsorized variances) to test for differences between groups.

Key words: Paired *t* test, independent *t* test, robust methods, Monte Carlo methods

Introduction

It is well known that the paired-samples *t* test has more power to detect a difference between the means of two groups as the correlation between the groups becomes larger. That is, as the population correlation coefficient, ρ, increases, the standard error of the difference between the means gets smaller, which in turn increases the magnitude of the *t* statistic (Kirk, 1999). Equation 1, the population variance of the difference between mean values, demonstrates how the standard error of the difference between the means ($\sigma_{\bar{X}_1-\bar{X}_2}$) is reduced as the value of ρ increases.

Katherine Fradette (umfradet@cc.umanitoba.ca) is a graduate student in the Department of Psychology. H. J. Keselman is a Professor of Psychology, email: famoy1kf@cmich.edu. Lisa Lix (lisa_lix@cpe.umanitoba.ca) is an Assistant Professor of Community Health Sciences. James Algina (algina@ufl.edu) is a Professor of Educational Psychology. Rand R. Wilcox (rwilcox@usc.edu) is a Professor of Psychology.

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} - 2\rho\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}, \qquad (1)$$

where $\sigma^2_{\bar{X}_j} = \sigma^2_j / n_j$ is the population variance of the mean for group j ($j = 1, 2$).

It must be kept in mind, however, that the independent-samples *t* test has twice the degrees of freedom of the paired-samples *t* test. Generally, an increase in degrees of freedom is accompanied by an increase in power. Thus, considering the loss of degrees of freedom for the paired-samples test, there is the question of just how large ρ must be in order for the paired-samples *t* test to achieve more power than the independent-samples *t* test.

Vonesh (1983) demonstrated that the paired-samples *t* test is more powerful than the independent-samples test when the correlation between the groups is .25 or larger. Furthermore, Zimmerman (1997) observed that many authors recommend the paired-samples *t* test only if "the two groups are highly correlated" and recommend the independent samples test if "they are uncorrelated or only slightly correlated" (p. 350). Zimmerman argued, however, that such authors often fail to take into account an important consequence of the use of the independent *t* test on dependent

observations. Namely, Zimmerman (1997) noted that the independence assumption is violated when the independent-samples *t* test is performed on groups that are correlated, even to a very small degree, and such a violation of the independence assumption distorts both Type I and Type II error rates.

Zimmerman (1997) compared the Type I error and power performance of the paired and independent-samples *t* tests for normally distributed data, varying the magnitude of ρ. He found that a correlation as small as .1 seriously distorted Type I error rates of the independent-samples *t* test. Thus, according to Zimmerman, the practice of employing the independent-samples *t* test when groups are slightly correlated fails to protect against distortion of the significance level and concluded that "a correlation coefficient of .10 or .15 is *not* sufficient evidence of independence, not even for relatively small sample sizes" (p. 359). Zimmerman also demonstrated an example in which, even when the correlation between two groups was as low as .1, the paired *t* test was more powerful than the independent-samples *t* test. Consequently, contrary to the recommendations of the authors he cites (e. g., Edwards, 1979; Hays, 1988; Kurtz, 1965), Zimmerman advocates the use of the paired-samples *t* test even when groups are only correlated to a very small degree (i.e., .1), when distributions are normal.

The question regarding how large ρ should be in order for the paired-samples *t* test to achieve more power than the independent-samples *t* test, when data are not normally distributed has not been examined (Wilcox, 2002). Evaluating the performance of statistics under nonnormality is important, given that psychological data are often not normal in shape (Micceri, 1989; Wilcox, 1990). Hence, the goal of this study was to extend Zimmerman's (1997) work by examining the Type I error and power rates of both the paired-samples and the independent-samples *t* tests when distributions were nonnormal, again varying the magnitude of ρ.

An investigation of the performance of both the paired and independent-samples *t* tests under nonnormality raises a problem, however. Both tests assume normally distributed data in the population. Violation of the normality assumption leads to distortion of Type I error rates and can lead to a loss of power to detect a difference between the means (MacDonald, 1999; Wilcox, 1997). Thus, in addition to an examination of the performance of the conventional (least squares) versions of the paired and independent-samples *t* tests, the performance of a robust version of each of the tests was also investigated.

The robust versions of the paired and independent-samples *t* tests involve substituting robust measures of location and scale for their least squares counterparts. Specifically, the robust versions of the tests substitute trimmed means for least squares means, and Winsorized variances for least squares variances. Calculation of the trimmed mean, which is defined later in Equation 7, involves trimming a specified percentage of the observations from each tail of the distribution (for symmetric trimming), and then computing the average of the remaining observations. The Winsorized variance, which is defined later in Equation 8, is computed by first Winsorizing the observations (see Equation 5), which also involves removing the specified percentage of observations from each end of the distribution. However, in this case the eliminated observations are replaced with the smallest and largest observation not removed from the left and right side of the distribution, respectively. The Winsorized variance is then computed in the same manner as the conventional least squares variance, using the set of Winsorized observations.

Numerous studies have shown that, under nonnormality, replacing least squares means and variances with trimmed means and Winsorized variances leads to improved Type I error control and power rates for independent groups designs (e.g., Keselman, Kowalchuk & Lix, 1998; Keselman, Wilcox, Kowalchuck & Olejnik, 2002; Lix & Keselman, 1998; Yuen, 1974), as well as dependent groups designs (e.g., Keselman, Kowalchuk, Algina, Lix & Wilcox, 2000; Wilcox, 1993). In particular, Yuen (1974) was the first to propose that trimmed means and Winsorized variances be used with Welch's (1938) heteroscedastic statistic in order to test for differences between two independent groups, when distributions are nonnormal and variances

are unequal. Thus, Yuen's method helps to protect against the consequences of violating the normality assumption and is designed to be robust to variance heterogeneity. Yuen's method reduces to Welch's (1938) heteroscedastic method when the percentage of trimming is zero (Wilcox, 2002). Yuen's method can also be extended to dependent groups.

It is important to note that while the conventional paired and independent-samples *t* statistics are used to test the hypothesis that the population means are equal ($H_0: \mu_1 = \mu_2$), the robust versions of the tests examine the hypothesis that the population trimmed means are equal ($H_0: \mu_{t1} = \mu_{t2}$). Although the robust versions of the procedures are not testing precisely the same hypotheses as their conventional counterparts, both the robust and conventional versions test the hypothesis that measures of the typical score are equal. In fact, according to many researchers, the trimmed mean is a better measure of the typical score than the least squares mean, when distributions are skewed (e.g., Keselman et al., 2002).

This study compared (a) the conventional (i.e., least squares means and variances) paired-samples *t* test, (b) the conventional independent-samples *t* test, (c) the robust (trimmed means and Winsorized variances) paired-samples *t* test, and (d) the robust independent-samples *t* test, based on their empirical rates of Type I error and power. As in Zimmerman's (1997) study with normal data, it was expected that as the size of the correlation between the groups increased, both the conventional and robust versions of the paired-samples *t* tests would perform better than their independent-samples counterparts, in terms of their ability to maximize power while maintaining empirical Type I error rates close to the nominal α level. It was also expected, based on previous findings (e.g., Keselman, et al., 1998; Keselman, et al., 2000; Keselman et al., 2002; Lix et al., 1998; Wilcox, 1993; Yuen, 1974), that the robust versions of both the paired and independent-samples *t* tests would perform better in terms of Type I error and power rates than the corresponding conventional versions.

Methodology

Definition of the Test Statistics

Conventional Methods

Suppose that n_j observations, $X_{1j}, X_{2j}, \ldots, X_{n_j j}$, are sampled from population j ($j = 1, 2$). In order to compute the conventional independent-samples *t* test, let $\bar{X}_j = \sum_i X_{ij} / n_j$ be the j^{th} sample mean ($i = 1, \ldots, n_j$; $N = \sum_j n_j$). Also let $S_j^2 = \sum_i \left(X_{ij} - \bar{X}_j\right)^2 / \left(n_j - 1\right)$ be the j^{th} sample variance. The estimate of the common (i.e., pooled) variance is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \qquad (2)$$

The test statistic for the conventional independent-samples *t* test is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad (3)$$

which is distributed as a *t* variable with $\nu = n_1 + n_2 - 2$ degrees of freedom, assuming normality and homogeneity of variances.

In order to compute the conventional paired-samples *t* test, which assumes that the two groups are dependent, let $S_{\bar{X}_j}^2 = S_j^2 / n_j$, where $S_{\bar{X}_j}$ is the estimate of the standard error of the mean of group *j*. An estimate of the correlation between the two groups is also needed to compute the paired-samples *t* statistic. The correlation is defined as $r = S_{12} / S_1 S_2$, where

$$S_{12} = \sum_i (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) / (n - 1),$$

and *n* represents the total number of pairs. The paired-samples test statistic is

$$T_{(PAIRED)} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2 - 2rS_{\bar{X}_1}S_{\bar{X}_2}}}, \qquad (4)$$

which is distributed as a *t* variable with $\nu = n - 1$

degrees of freedom, assuming normality.

Robust Methods

Suppose, again, that n_j observations, $X_{1j}, X_{2j}, \ldots, X_{n_j j}$, are sampled from population *j*. For both the independent-samples and paired-samples *t* tests, first let $X_{(1)j} \le X_{(2)j} \le \cdots \le X_{(n_j)j}$ be the ordered observations of group *j*, and let γ be the percentage of observations that are to be trimmed from each tail of the distribution. Also let $g_j = [\gamma n_j]$, where [*x*] is the largest integer $\le x$. To calculate the robust versions of both statistics we must first Winsorize the observations by letting

$$\begin{aligned} Y_{ij} &= X_{(g_j+1)j} \text{ if } X_{ij} \le X_{(g_j+1)j} \\ &= X_{ij} \text{ if } X_{(g_j+1)j} < X_{ij} < X_{(n_j-g_j)j} \\ &= X_{(n_j-g_j)j} \text{ if } X_{ij} \ge X_{(n_j-g_j)j} \end{aligned} . \quad (5)$$

The sample Winsorized mean is defined as

$$\bar{Y}_{Wj} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} . \quad (6)$$

The sample trimmed mean for the j^{th} group is also required to compute the robust versions of the paired and independent-samples *t* tests and is defined as

$$\bar{X}_{tj} = \frac{1}{h_j}\sum_{i=g_j+1}^{n_j-g_j} X_{(i)j} , \quad (7)$$

where $h_j = n_j - 2g_j$. The sample Winsorized variance for the robust independent-samples *t* test is

$$S^2_{Wj} = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(Y_{ij} - \bar{Y}_{Wj}\right)^2 , \quad (8)$$

where Y_{ij} and $\bar{Y}_{Wj}$ are defined in Equations 5 and 6, respectively. Finally, let

$$d_j = \frac{\left(n_j - 1\right)S^2_{Wj}}{h_j\left(h_j - 1\right)} . \quad (9)$$

Then the robust independent-samples *t* test is

$$T_Y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2}} , \quad (10)$$

which is approximately distributed as a *t* variable with degrees of freedom

$$\nu_Y = \frac{\left(d_1 + d_2\right)^2}{d_1^2/\left(h_1 - 1\right) + d_2^2/\left(h_2 - 1\right)} . \quad (11)$$

To compute the robust paired-samples *t* test, as enumerated by Wilcox (2002), the paired observations must first be Winsorized, as in Equation 5. It is important to note that when Winsorizing the observations for the paired-samples *t* statistic, care must be taken to maintain the original pairing of the observations. The sample size for the robust version of the paired-samples *t* test is $h = n - 2g$, where *n* is the total number of pairs. Let

$$d_j = \frac{1}{h\left(h - 1\right)}\Sigma_i\left(Y_{ij} - \bar{Y}_{Wj}\right)^2 , \quad (12)$$

and

$$d_{12} = \frac{1}{h\left(h - 1\right)}\Sigma_i\left(Y_{i1} - \bar{Y}_{W1}\right)\left(Y_{i2} - \bar{Y}_{W2}\right), \quad (13)$$

where Y_{ij} and $\bar{Y}_{Wj}$ are defined in Equations 6 and 7, respectively. The test statistic for the robust paired-samples *t* test is

$$T_{Y(PAIRED)} = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2 - 2d_{12}}} , \quad (14)$$

which is approximately distributed as a *t* variable with $\nu = h - 1$ degrees of freedom.

Simulation Procedures

Empirical Type I error and power rates were collected for the conventional and robust versions of the paired and independent-samples *t* tests using a Monte Carlo procedure. Thus, a total of four tests were investigated: (a) the conventional paired-samples *t* test, (b) the

conventional independent-samples *t* test, (c) the robust paired-samples *t* test, and (d) the robust independent-samples *t* test. Two-tailed tests were performed on each of the four procedures.

Four variables were manipulated in the study: (a) sample size, (b) magnitude of the population correlation coefficient, (c) magnitude of the difference between groups, and (d) population distribution. Following Zimmerman (1997), four sample sizes (*N*) were investigated: 10, 20, 40, and 80, and population correlations (ρ) ranging from -.5 to .5, in increments of .1, were induced.

The difference in the mean (trimmed mean) value for the two populations was also manipulated. When empirical Type I error rates were investigated, there was no difference between the groups. When empirical power rates were investigated, three values of the effect size were investigated; the difference between the groups was set at .25, .5, and .75. These values were chosen in order to avoid ceiling and floor effects, a practice that has been employed in other studies (e.g., Keselman, Wilcox, Algina, Fradette, & Othman, 2003).

There were two population distribution conditions. Data for both groups were generated either from an exponential distribution or a chi-squared distribution with one degree of freedom (χ_1^2). Skewness and kurtosis values for the exponential distribution are $\gamma_1 = 2$ and $\gamma_2 = 6$, respectively. Skewness and kurtosis values for the χ_1^2 distribution are $\gamma_1 = \sqrt{8}$ and $\gamma_2 = 12$, respectively.

For the robust versions of both the paired and the independent-samples *t* tests, the percentage of trimming was 20%; thus, 20% of the observations from each tail of the distribution were removed. This proportion of trimming was chosen because it has been used in other studies (e.g., Keselman et al., 1998; Keselman et al., 2000; Keselman et al., 2002; Lix et al., 1998) and because 20% trimming has previously been recommended (e.g., Wilcox, 1997).

In order to generate the data for each condition, the method outlined in Headrick and Sawilowsky (1999) for generating correlated multivariate nonnormal distributions was used. First, the SAS generator RANNOR (SAS Institute, 1989) was used to generate pseudo-random normal variates, Z_i ($i = 1, \ldots, N$). Next, the Z_i s were modified using the algorithm

$$Y_{ij} = rZ_i + \sqrt{1-r}E_{ij}, \qquad (15)$$

where the E_{ij} s are pseudo-random normal variates. In the case of this study, the E_{ij} s were also generated by the SAS generator RANNOR. The variable r is determined as in Headrick and Sawilowsky (1999), and is dependent on the final desired population correlation (ρ). Both Y_{i1} and Y_{i2} are random normal deviates with a correlation of r^2. Finally, the Y_{ij} s generated for the study were further modified in order to obtain nonnormally distributed observations, via the algorithm

$$Y_{ij}^* = a + bY_{ij} + (-a)Y_{ij}^2 + dY_{ij}^3, \qquad (16)$$

where *a*, *b*, and *d* are constants that depend on the desired values of skewness (γ_1) and kurtosis (γ_2) of the distribution, and can be determined by solving equations found in Fleishman (1978, p. 523). The resultant Y_{ij}^* s are nonnormal deviates with zero means and unit variances, and are correlated to the desired level of ρ, which is specified when determining *r*.

Observations with mean μ_j (or μ_{tj}) and variance σ_j^2 were obtained via $X_{ij} = \mu_j + \sigma_j \times Y_{ij}^*$. The means (trimmed means) varied depending on the desired magnitude of the difference between the two groups. In order to achieve the desired difference, constants were added to the observations in each group. The value of the constants, corresponding to each of the four difference conditions investigated, were (a) 0, 0, (b) .25, 0, (c) .5, 0, and (d) .75, 0. These values were added to each observation in the first and second group, respectively. Thus, μ_j (μ_{tj}) represents the value of the constants corresponding to a given desired difference. Variances were set to $\sigma_j^2 = 1$ in all conditions. When using trimmed means, the empirically

determined population trimmed mean μ_t was subtracted from the Y_{ij}^{*} variates before multiplying by σ_j (see Keselman et al., 2002 for further discussion regarding the generation of variates to be used with trimming). Ten thousand replications of the data generation procedure were performed for each of the conditions studied.

Results

Type I Error Rates

Each of the four investigated tests was evaluated based on its ability to control Type I errors, under conditions of nonnormality. In the case of the two versions of the independent-samples *t* tests, the independence assumption was also violated when ρ was not equal to zero. In order for a test to be considered robust, its empirical rate of Type I error ($\hat{\alpha}$) had to be contained within Bradley's (1978) liberal criterion of robustness: $0.5\alpha \le \hat{\alpha} \le 1.5\alpha$. Hence, for this study, in which a five percent nominal significance level was employed, a test was considered robust in a particular condition if its empirical rate of Type I error fell within the $.025 - .075$ interval. A test was considered to be nonrobust in a particular condition if $\hat{\alpha}$ fell outside of this interval. Tables 1 and 2 display the range of Type I errors made by each of the investigated tests across all samples sizes (*N* = 10, 20, 40, 80), as a function of ρ. We felt it was acceptable to enumerate a range across all sample sizes investigated because at all values of *N*, a similar pattern of results was observed.

Table 1: Range of Proportion of Type I Errors for All Tests Under the Exponential Distribution

Exponential Distribution				
Rho (ρ)	Conventional Procedure		Robust Procedure	
	Independent	Paired	Independent	Paired
-0.5	.116 - .143	.060 - .093	.103 - .108	.051 - .057
-0.4	.100 - .128	.056 - .085	.092 - .099	.052 - .054
-0.3	.089 - .116	.055 - .083	.078 - .092	.047 - .054
-0.2	.081 - .108	.059 - .086	.070 - .080	.049 - .057
-0.1	.071 - .091	.059 - .078	.062 - .067	.049 - .053
0	.042 - .048	.039 - .049	.038 - .046	.035 - .045
0.1	.035 - .043	.042 - .053	.031 - .038	.031 - .049
0.2	.025 - .029	.044 - .050	.024 - .031	.030 - .052
0.3	.019 - .021	.042 - .053	.017 - .021	.028 - .048
0.4	.011 - .012	.039 - .052	.012 - .016	.03 - .044
0.5	.006 - .007	.04 - .047	.006 - .01	.028 - .045

Table 2: Range of Proportion of Type I Errors for All Tests Under the χ_1^2 Distribution

Chi-Squared Distribution (χ_1^2)				
Rho (ρ)	Conventional Procedure		Robust Procedure	
	Independent	Paired	Independent	Paired
-0.5	.120 - .171	.068 - .129	.102 - .107	.056 - .073
-0.4	.100 - .161	.060 - .125	.090 - .096	.056 - .068
-0.3	.093 - .145	.063 - .118	.079 - .089	.051 - .066
-0.2	.087 - .135	.067 - .114	.070 - .082	.052 - .067
-0.1	.075 - .114	.064 - .102	.063 - .068	.052 - .058
0	.038 - .046	.034 - .046	.026 - .045	.025 - .042
0.1	.031 - .041	.033 - .049	.023 - .036	.022 - .042
0.2	.026 - .029	.035 - .046	.020 - .030	.023 - .044
0.3	.020 - .021	.033 - .052	.018 - .023	.023 - .043
0.4	.011 - .015	.035 - .051	.015 - .018	.022 - .046
0.5	.006 - .011	.035 - .045	.009 - .013	.020 - .042

Table 1 displays the range of empirical Type I error rates for each test, as a function of ρ, under the exponential distribution condition. It is apparent from the table that both versions of the paired-samples *t* test maintained Type I errors near the nominal level of significance, α. In fact, only 6 of 44 values fell outside the range of Bradley's $.025 - .075$ interval for the conventional paired *t* test; none did for the robust paired *t* test. Thus, for data that follow an exponential distribution, the robust paired *t* test was insensitive to nonnormality at every value of ρ. A comparison of the conventional and robust versions of the paired *t* test in Table 1 reveals that, in particular, the robust version was more effective at controlling Type I errors when the population correlation (ρ) between the groups was negative.

Table 1 also shows that the independent-samples tests were not as robust, overall, as their paired-samples counterparts. In fact, the total number of values that fell outside of the range of Bradley's liberal criterion was 30 and 26 (out of 44) for the conventional and robust versions of the independent *t* test, respectively. Thus, the robust independent *t* test was indeed slightly more robust, overall, than the conventional independent *t* test. Both versions of the independent-samples *t* test were effective at controlling Type I errors when the population correlation (ρ) was zero; however, this control was reduced the more that ρ deviated from zero.

An inspection of Table 2, which displays the range of Type I errors for the tests for the χ_1^2 distribution, reveals a pattern of results similar to that for the exponential distribution. However, all of the tests were somewhat less robust under the χ_1^2 distribution than the exponential distribution condition. That

is, nonrobust liberal values were greater in value for χ_1^2 data than for exponentially distributed data. Specifically, the total number of values that fell outside of Bradley's liberal interval for the conventional versions of the paired and independent-samples *t* tests were 12 and 31 (out of 44), respectively. The total number of nonrobust values for the robust versions of the paired and independent-samples *t* tests were five and 28, respectively.

Power Rates

The four tests were also evaluated based on empirical power rates. Therefore, each test was judged on its ability to detect a true difference between the trimmed means of the groups (in the case of the robust tests), or the least squares means of the groups (in the case of the conventional tests). Figures 1, 2, and 3 display the power of each of the investigated tests to detect a true difference between the (trimmed) means of the groups, as a function of the magnitude of the difference between the (trimmed) means. The results portrayed in these figures were averaged over all sample sizes. While the power rates of the tests increased as the size of *N* increased, again, we felt it was acceptable to collapse over the sample size conditions because the tests showed a similar pattern of results in relation to one another for all values of *N*.

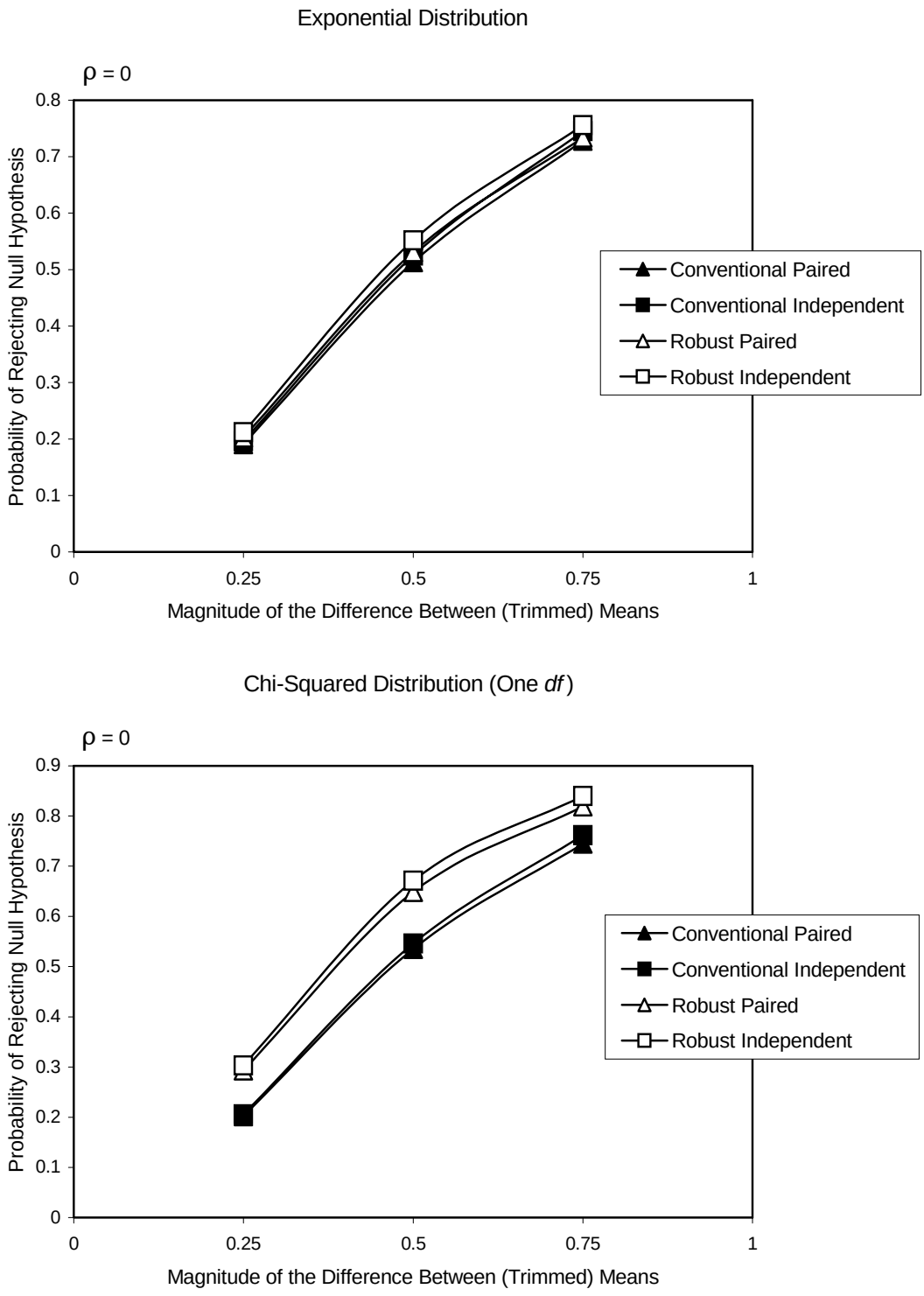

Figure 1. Probability of rejecting H_0 for the conventional and robust paired and independent-samples *t* tests; $\rho = 0$.

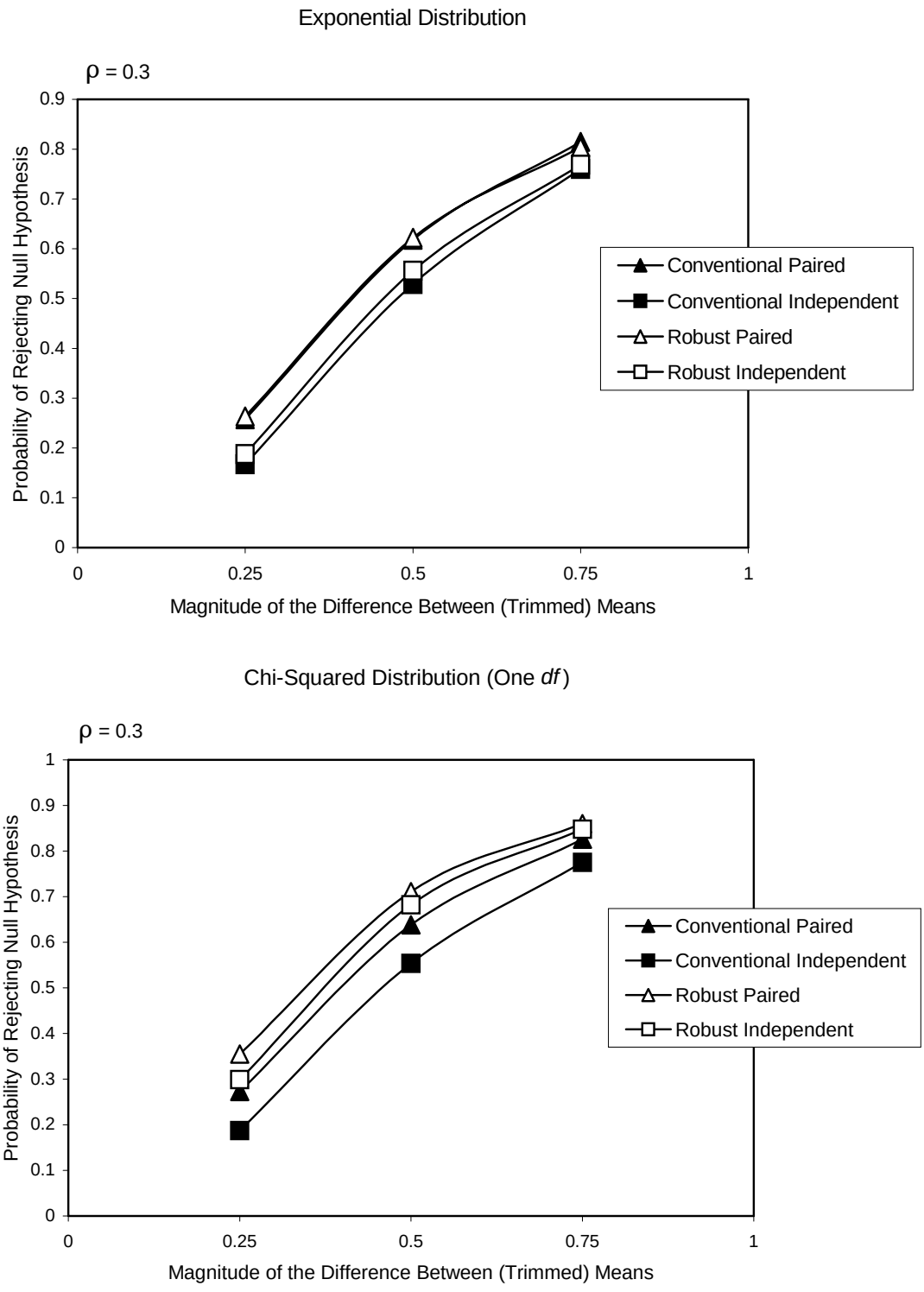

Figure 2. Probability of rejecting H_0 for the conventional and robust paired and independent-samples *t* tests; $\rho = 0.3$.

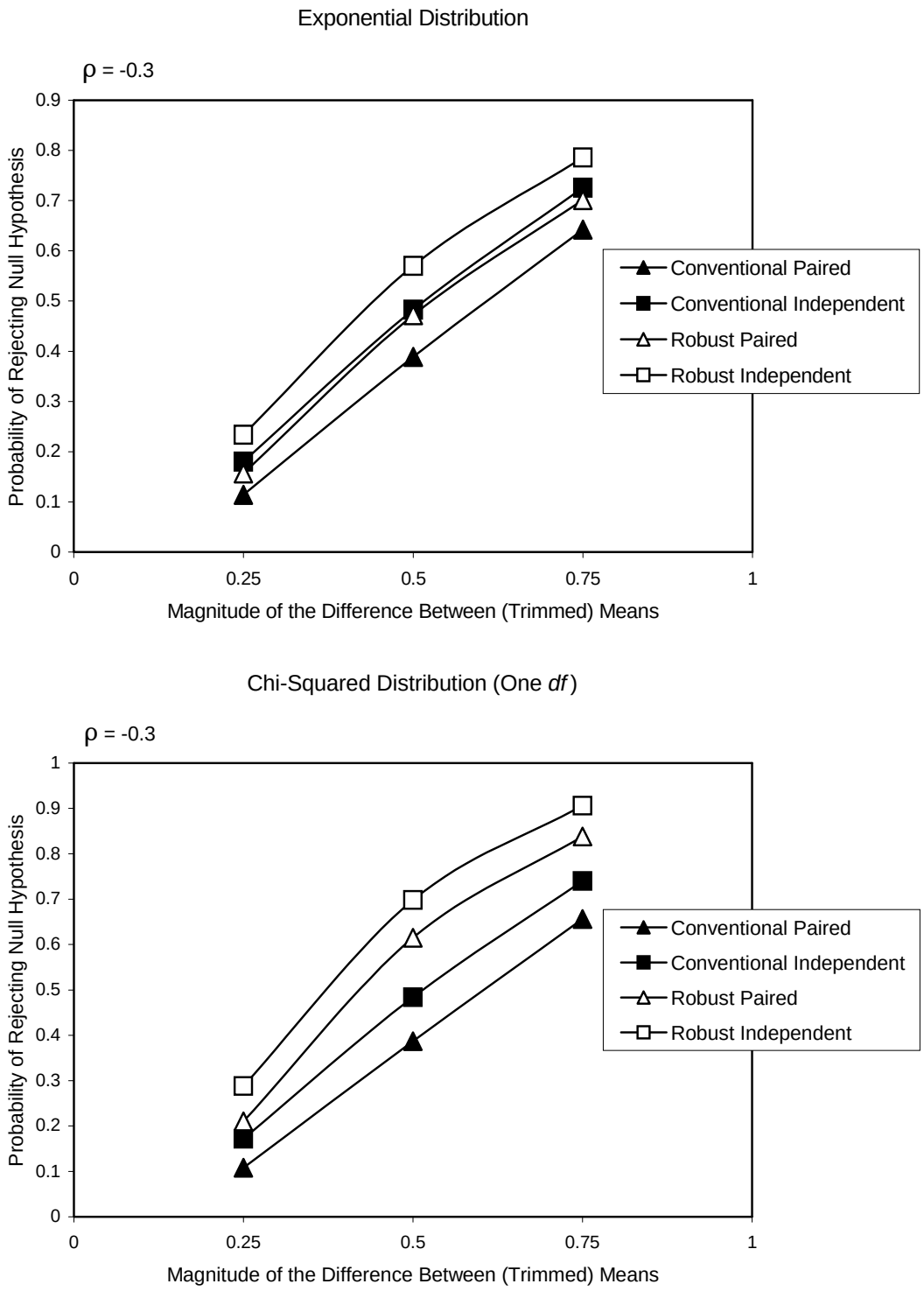

Figure 3. Probability of rejecting H_0 for the conventional and robust paired and independent-samples *t* tests; $\rho = -0.3$.

Figure 1 displays the power rates of the tests for both the χ_1^2 and the exponential distributions when $\rho = 0$. The upper portion of the figure reveals that when data followed an exponential distribution, the power functions of the four tests were quite similar, with the empirical power of the robust versions only slightly higher than the corresponding power of the conventional versions. However, an inspection of the lower portion of Figure 1 indicates that under the χ_1^2 distribution, the power functions of the robust tests were considerably higher than those of both conventional versions. In addition, Figure 1 shows that when no correlation existed between the groups, the power functions of the independent-samples *t* tests were slightly higher than their paired-samples counterparts.

Figure 2 shows the power functions of the tests for both the χ_1^2 and exponential distributions when $\rho = .3$. The upper portion of Figure 2 indicates that when the data were exponentially distributed and positively correlated, the power functions of both versions of the paired-samples *t* test were higher than those of the independent-samples tests. The lower portion of the figure, which displays power for the χ_1^2 distribution for this same value of ρ, demonstrates that while the power function of each of the paired-samples *t* tests was higher than its respective independent-samples counterpart, the power rates of both robust tests were higher than those of the conventional tests.

Figure 3 displays the power rates of the tests for the χ_1^2 and exponential distributions when $\rho = -.3$. Unlike the results obtained for positively correlated data, the paired-samples *t* tests showed no apparent power advantage over the independent-samples *t* tests when the groups were negatively correlated, for either the exponential or the χ_1^2 distributions. In fact, the figure shows that the power functions of the independent-samples *t* tests were higher than their paired-samples counterparts under both distributions. The lower portion of Figure 3 shows that under the χ_1^2 distribution, while the power functions of both versions of the independent-samples *t* test were higher than their corresponding versions of the paired-samples test, the power rates of both robust tests were higher than the conventional tests, as was the case with the other levels of ρ.

Conclusion

Four different statistics for testing the difference between two groups were investigated based on their power to detect a true difference between two groups and their ability to control Type I errors. The primary objective for conducting the study was to determine which of the tests would perform best when the data for the two groups were correlated and the assumption of a normal distribution of the responses was violated.

Although empirical Type I error and power rates are two separate measures of a test's effectiveness, in order to evaluate the overall performance of the investigated procedures, power and Type I error rates must be considered concomitantly. The reason for this is that if a test does not maintain the rate of Type I errors at or around the nominal α level, this can cause a distortion in power. Figures 4 and 5 provide a summary of the results for the exponential and χ_1^2 distributions, respectively. These figures were included to allow the reader to easily examine the Type I error and power rates of each of the distributions concurrently.

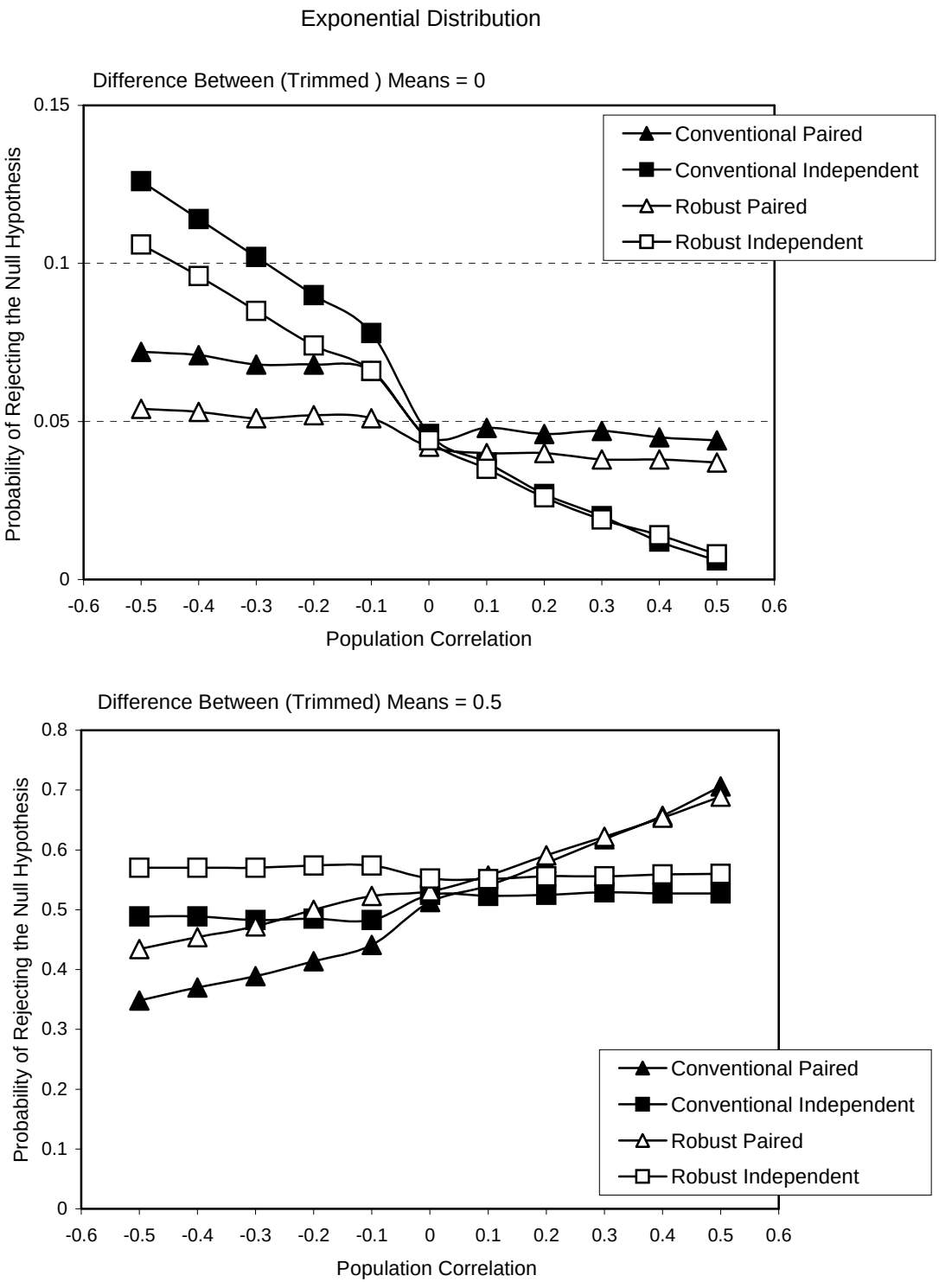

Figure 4. Probability of rejecting H_0 as a function of ρ and the magnitude of the difference between (trimmed) means for the conventional and robust paired and independent-samples *t* tests exponential distribution.

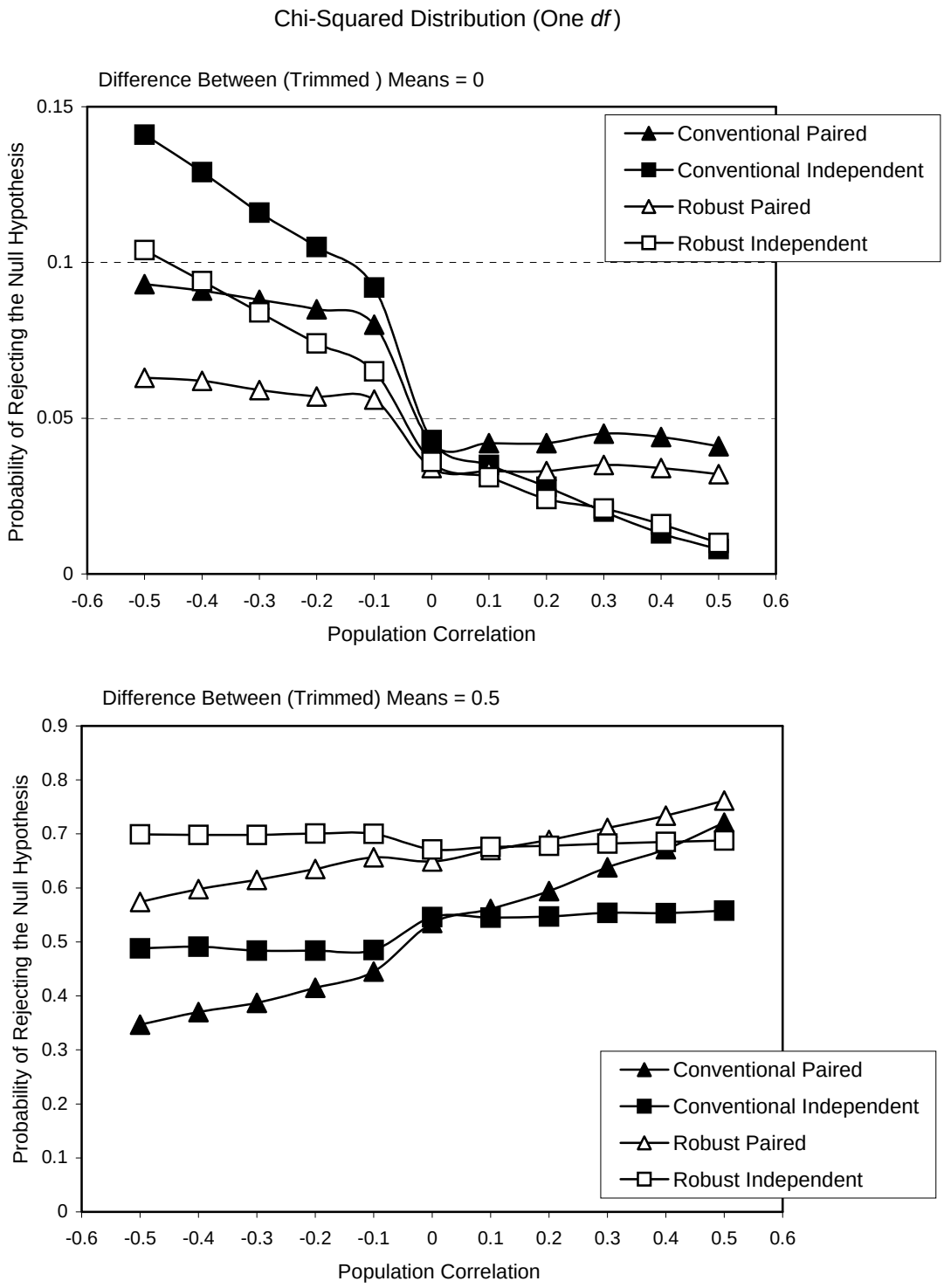

Figure 5. Probability of rejecting H_0 as a function of ρ and the magnitude of the difference between (trimmed) means for the conventional and robust paired and independent-samples *t* tests under the χ_1^2 distribution.

As the results indicated, the only time the independent tests maintained the Type I error rate close to the nominal level was when there was no correlation between the groups; this ability grew worse as ρ got larger. In fact, the Type I error control of the independent *t* tests began to break down when the correlation between the groups was as small as $\pm.1$. Thus, with the exception of the $\rho = 0$ condition, both the robust and the conventional versions of the independent *t* test were quite poor at controlling Type I errors. Because of this distortion of the Type I error rate, the powers of the independent tests are not interpretable (Zimmerman, 1997) when ρ is not equal to zero.

Both versions of the paired *t* test, however, did a much better job of controlling Type I errors than their independent-samples counterparts when there was a correlation between the groups, for nonnormal data. Because the paired-samples *t* tests maintained Type I errors close to the nominal level, the empirical power rates of the paired *t* tests, unlike those of the independent tests, can be taken to accurately represent their ability to detect a true difference between the groups. Thus, as expected, when power and Type I error rates are both taken into account, it can be said that the paired *t* tests were more effective than their independent samples counterparts when groups were correlated, even when this correlation was low (i.e., $\pm.1$). This finding agrees with Zimmerman's (1997) results for normally distributed data.

Furthermore, the robust paired-samples *t* test was more effective, in terms of Type I error control, than the conventional paired test. The robust paired test was also consistently more powerful than the conventional version, and this power advantage increased as skewness and kurtosis in the population increased. Therefore, as expected, the robust version of the paired-samples *t* test performed better than the conventional version of the test, for nonnormal data. This result is supported by many other studies involving trimmed means and Winsorized variances (e.g., Keselman, et al., 1998; Keselman, et al., 2000; Keselman et al., 2002; Lix et al., 1998; Wilcox, 1993; Yuen, 1974).

In conclusion, there need only be a small positive or negative correlation between two groups in order for the paired *t* test to be more effective than the independent *t* test when the data are nonnormal. In fact, although Vonesh (1983) showed that there needs to be a correlation of at least .25 in the population for the paired *t* test to be more powerful than the independent test, when the distortion of Type I error rates, resulting from the application of the independent-samples *t* test on dependent data, was taken into account, the paired-samples *t* tests performed best when the correlation was as low as $\pm.1$. Thus, just as Zimmerman (1997) cautions when dealing with normal data, researchers should take care to ensure that their data is not correlated in any way when using the independent *t* test on nonnormal data, lest the existence of even a slight dependence alters the significance level of the test. In addition, given that the population distributions were not normal in shape, the robust version of the paired *t* test performed the best under all the conditions that were studied. Thus, based on the results of this investigation, it is recommended that researchers use the robust paired-samples *t* test, which employs trimmed means and Winsorized variances, when dealing with nonnormal data.

References

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

Edwards, A. L. (1979). *Multiple regression and the analysis of variance and covariance.* New York: Freeman.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43,* 521-532.

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

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika, 64,* 25-35.

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology, 53,* 175-191.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. *Psychometrika, 63,* 145-163.

Keselman, H. J., Wilcox, R. R., Algina, J., Fradette, K., Othman, A. R. (2003). *A power comparison of robust test statistics based on adaptive estimators.* Submitted for publication.

Keselman, H. J., Wilcox, R. R., Kowalchuk, R. K., & Olejnik, S. (2002). Comparing trimmed or least squares means of two independent skewed populations. *Biometrical Journal, 44,* 478-489.

Kirk, R. E. (1999). *Statistics: An introduction.* Orlando, FL: Harcourt Brace.

Kurtz, K. H. (1965). Foundations of psychological research. Boston: Allyn & Bacon.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement, 58,* 409-429.

MacDonald, P. (1999). Power, Type I error, and Type III error rates of parametric and nonparametric statistical tests. *Journal of Experimental Education, 67,* 367-380.

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

SAS Institute Inc. (1989). SAS/IML software: Usage and reference, version 6 (1st ed.). Cary NC: Author.

Vonesh, E. F. (1983). Efficiency of repeated measures designs versus completely randomized designs based on multiple comparisons. *Communications in Statistics A: Theory and Methods, 12,* 289-301.

Welch, B. L. (1938). 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 two independent groups. *Biometrical Journal, 36,* 259-273.

Wilcox, R. R. (1993). Analysing repeated measures or randomized block designs using trimmed means. *British Journal of Mathematical and Statistical Psychology, 46,* 63-76.

Wilcox, R. R. (1997). *Introduction to robust estimation and testing.* San Diego: Academic Press.

Wilcox, R. R. (2002). *Applying contemporary statistical methods.* San Diego: Academic Press.

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika, 61,* 165-170.

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