Psychological Test and Assessment Modeling, Volume 55, 2013 (1), 39-61

Robustness and power of the parametric *t* test and the nonparametric Wilcoxon test under non-independence of observations

Wolfgang Wiedermann[1] *& Alexander von Eye*[2]

Abstract

A large part of previous work dealt with the robustness of parametric significance tests against non-normality, heteroscedasticity, or a combination of both. The behavior of tests under violations of the independence assumption received comparatively less attention. Therefore, in applications, researches may overlook that robustness and power properties of tests can vary with the sign and the magnitude of the correlation between samples. The common paired *t* test is known to be less powerful in cases of negative between-group correlations. In this case, Bortz and Schuster (2010) recommend the application of the nonparametric Wilcoxon test. Using Monte-Carlo simulations, we analyzed the behavior of the *t*- and the Wilcoxon tests for the one- and two-sample problem under various degrees of positive and negative correlations, population distributions, sample sizes, and true differences in location. It is shown that already minimal departures from independence heavily affect Type I error rates of the two-sample tests. In addition, results for the one-sample tests clearly suggest that the sign of the underlying correlation cannot be used as a basis to decide whether to use the *t* test or the Wilcoxon test. Both tests show a dramatic power loss when samples are negatively correlated. Finally, in these cases, the well-known power advantage of the Wilcoxon test diminishes when distributions are skewed and samples are small.

Key words: robustness, power, independence assumption, *t* test, Wilcoxon test

[1] *Correspondence concerning this article should be addressed to:* Wolfgang Wiedermann, PhD, University of Vienna, Unit of Research Methods, Liebiggasse 5, A-1010 Vienna, Austria; email: wolfgang.wiedermann@univie.ac.at

[2] University of Vienna, Department of Psychology

Ever since the work of W. S. Gosset (‘Student’, 1908) and R. A. Fisher (1925) on statistical inference about differences in means (*Student’s t test*), a good deal of research focused on the properties of the *t* statistic. When the assumptions of normality, homoscedasticity, and independence of observations are met, Student’s two-sample *t* test was shown to be the optimal procedure for the comparison of means from independent samples (Hodges & Lehmann, 1956; Randles & Wolfe, 1979). However, in empirical data, violations of one or more assumptions might exist, and robustness properties of significance tests are of great interest. Early theoretical findings suggest that the two-sample *t* test is fairly robust against violations of the normality assumption (e.g., Bartlett, 1935). This result was confirmed in numerous simulation studies (e.g., Borneau, 1960; Neave & Granger, 1968; Posten, 1978, 1984; Rasch & Guiard, 2004). Although the two-sample *t* test is able to protect the nominal significance level α under non-normality, considerable evidence exists that the nonparametric Wilcoxon-Mann-Whitney *U* test is robust and even more powerful under various non-normal distributions (Hodges & Lehmann, 1956; Neave & Granger, 1968; Randles & Wolfe, 1979). In addition, it has been demonstrated that the two-sample *t* test is robust against violations of equality of variances when sample sizes are equal (e.g., Hsu, 1938; Scheffè, 1970; Posten, Yeh & Owen, 1982, Tuchscherer & Pierer, 1985; Zimmerman, 2006). When both, variances and sample sizes are unequal, the probability of the Type I error exceeds the nominal significance level if the larger variance is associated with the smaller sample size, and vice versa (Moder, 2010; Wiedermann & Alexandrowicz, 2007; Zimmerman, 2006). In this case, Welch’s *t* test (Welch, 1938, 1947) is recommended as an adequate alternative (see also a recent reminder of Rasch, Kubinger & Moder, 2011).

Although it is well known that the two-sample *t* test assumes independent observations, less attention has been paid to non-independence. Here, the distinction of between-group and within-group dependency has to be made. Between-group dependence refers to the fact that observations of two samples are correlated (for example, data obtained from a matched samples design or repeated observations). For the analysis of repeated measurements the term “one-sample problem” is commonly used, which underlines the fact that only one sample of research units is drawn from the underlying population of interest and the construct of interest is measured repeatedly (for details see Rasch, Kubinger & Yanagida, 2011). In contrast, within-group dependence means that scores are correlated with other scores within the same group (for example, if subjects influence each other’s responses). For both types of dependency, simulation studies showed that Type I error rates of the two-sample *t* test are strongly affected. It is important to note that the different types of dependencies can have different effects on the behavior of parametric significance tests. For positive between-group correlations (e.g., higher scores in a baseline assessment are associated with higher scores in a follow-up assessment), the probability of a Type I error falls below the nominal significance level. In contrast, positive within-group correlations (e.g., higher scores obtained in a subset of the sample are associated with higher scores in another subset of the same sample) increase the Type I error rates (Cochran, 1947; Lissitz & Chardos, 1975; Zimmerman, Williams & Zumbo, 1993; Zimmerman, 1997). Paired data (i.e., data exhibiting a non-zero between-group correlation) can easily be analyzed using statistical tests developed for the one-sample problem. Given that difference scores follow a normal distribution, the paired *t* test (essentially a one-sample *t* test performed on differences of sample values) is the optimal significance test (Hodges & Lehmann, 1956; Randles & Wolfe, 1979).

Several studies have shown that the paired *t* test is highly robust against violations of the normality assumption with respect to the Type I error (e.g., Herrendörfer, Rasch & Feige, 1983; Posten, 1979; Rasch & Guiard, 2004). However, for various non-normal densities, the nonparametric Wilcoxon-matched-pairs-signed-ranks test has proven to be robust and more powerful (Blair & Higgins, 1985). It is important to note that, although these test statistics were developed for paired data, the tests still assume independent observations within samples. Previous studies reported heavily biased Type I error rates in cases where observations systematically carry information about other observations (Chlaß & Krüger, 2007; Guiard & Rasch, 2004; von Eye, 1983, 2004). The present study focuses on between-group dependencies.

In practical data analysis, researchers may overlook that negative between-group correlations can have different effects on both two-sample and one-sample tests than positive between-group correlations. Only few studies analyzed the behavior of significance tests considering positive as well as negative correlations (for an exception see Zimmerman, 1997, 2012). This seems surprising considering that the occurrence of negative between-group correlations may be a common result of matching pairs. To give an example, Hays (1994) states that personality dominance of married couples might be negatively correlated, if highly dominant women tend to marry men with lower tendencies for dominance and vice versa.

In this article, we ask questions concerning the consequences of negative between-group correlations. Bortz and Schuster (2007) indicate that the paired *t* test is less powerful under negative correlations, which is in line with simulation results of Zimmerman (1997). The authors recommend using the nonparametric Wilcoxon test instead (Bortz & Schuster, 2007, p. 125). The current study aims to systematically investigate the behavior of the parametric *t*- and the nonparametric Wilcoxon tests, developed for the one- and the two-sample problems, under various degrees of positive and negative between-group correlation, various sample sizes, and various distributions. It will be shown that these tests perform virtually identically with respect to the power to detect true differences between samples.

Methods

In this article, we report results from a Monte-Carlo study in which we focus on dependencies between groups (e.g., matched pairs or repeated observations). To perform the simulations, a program was written using the R statistical environment (R Core Team, 2012) that varied four factors: Between-group correlation (ρ), type of distribution, sample size (n), and difference in means ($\mu_1 - \mu_2$). The following sections describe the factors in detail.

Between-group correlation. To mimic violations of the between-group independence assumption, correlations between two samples y_1 and y_2 of $\rho = -0.8, \ldots(0.1)\ldots, 0.8$ were induced by adding a multiple of one random variate to another. The multiplicative constant c was chosen to obtain the desired correlation. Let y_1, y_2, and z be independent standard normally distributed variables with zero means and unit variances. If $c = \sqrt{\frac{\rho}{(1-\rho)}}$, $x_1 = \frac{y_1 + cz}{\sqrt{1+c^2}}$ and $x_2 = \frac{y_2 + cz}{\sqrt{1+c^2}}$ follow a standard normal distribution with a between-group correlation of ρ (cf. Wiedermann & Alexandrowicz, 2011; Zimmerman, 2012). To generate negative correlations, first, the desired positive correlation ρ was invoked and then the scores of x_1 were multiplied by (–1).

Type of distribution. N(0,1) distributed variables were generated using the Ziggurat method of Marsaglia and Tsang (2000). Next, the normal variates were transformed to simulate various non-normal distributions. The following four non-normal shapes were realized (see, for example, Evans, Hastings & Peacock, 2000):

- *Uniform distribution:* Uniformly distributed variates were generated using a probability integral transformation: $u = F\ x\ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\ -x'^2/2\ dx'$, where x and F denote the standard normally distributed random variable and the cumulative distribution function of the standard normal distribution, respectively. Resulting variates are expected to exhibit a skewness of 0 and kurtosis of 1.8.
- *Logistic distribution:* Uniformly distributed variates (u) were transformed using $x = \log\ u/(1-u)$ and are expected to show a skewness of 0 and a kurtosis of 4.2.
- *Gumbel distribution:* Uniformly distributed variates were transformed using $x = -\log\ -\log u$. The resulting distribution is expected to show an elevated skewness and a kurtosis of 1.14 and 5.4, respectively.
- *Exponential distribution:* Exponentially distributed values were obtained applying $x = -\log\ u\ -1$ and are expected to show a skewness of 2 and a kurtosis of 9.

Differences in means. To analyze Type I error rates as well as the power of the significance tests, constants were added to one sample to produce the following differences in means: $\mu_1 - \mu_2$ = 0, 0.25, 0.50, 0.75.

Sample size. The number of observations (n) varied from 30 to 250, in increments of 20.

For each experimental cell of the 5 (distribution shapes) x 17 (correlation) x 12 (sample sizes) x 4 (differences in means) design 5000 repetitions were realized. In each repetition, the samples x_1 and x_2 were evaluated using the two-sample t test, the (standard normal-approximated) Wilcoxon-Mann-Whitney U test, the paired t test, and the (standard normal-approximated) Wilcoxon test. Thus, a total of 20,400,000 test statistics for each significance test were retained and further analyzed applying standard ANOVA techniques (for an example see, von Eye, 2004). All significance tests were performed non-directionally under a nominal significance level of α = 5%. To evaluate the robustness of Type I error rates, a 20% robustness criterion was chosen. Thus, for α = 5%, a significance test is considered robust if the empirical Type I error rates do not exceed the interval 4 – 6%. The results presented in Figures 2 – 4 were obtained using 50,000 repetitions. The implemented R program is freely available from the authors upon request.

Results

Type I error

Due to the large quantity of simulation outcomes, results of the Type I error simulation are only presented for $\rho = -0.8, \ldots (0.2), \ldots (0.8)$ and $n = 30, 150, 250$. Findings for the experimental conditions not shown here are very similar to the presented results and can be obtained from the authors upon request. Tables 1 and 2[3] show the Type I error rates of the four significance tests for the five distributions. Each entry gives the relative frequency of rejecting the null hypothesis. As expected, all tests are very well able to protect the nominal significance level of 5% for completely independent samples ($\rho = 0$), for all simulated sample sizes and distributions. For $\rho \neq 0$, the significance tests of the one-sample problem are able to keep Type I error rates close to the 5% level, also as expected. In contrast, the test statistics of the two-sample procedures are heavily biased in these cases.

Two different effects were observed depending on the sign of the population correlation. For negatively correlated samples, the probabilities of a Type I error are far above the nominal significance level for the two-sample *t* test and the *U* test. However, both tests become overly conservative for positively correlated samples (i.e., the Type I error rates fall far below 5%). Again, this holds for normal as well as for non-normal distributions. Figure 1 illustrates that the magnitude of these biases varies with the magnitude of the correlation. It can be seen that the percentiles of the two-sample *t* statistic are not independent of the level of population correlation which leads to biased decisions concerning the null hypothesis (Figure 1, left panel). In contrast, the *t* values of the paired *t* test are unaffected by the degree of correlation (Figure 1, right panel).

To investigate the sensitivity of the simulated test statistics to the factors of the simulation, we employed ANOVAs. Table 3 summarizes the ANOVA results using the corresponding *t* and *z* values as dependent variables. Due to computational limits, the ANOVAs were restricted to $n = 30, \ldots, 230$ in increments of 40 and $\rho = -0.8, \ldots, 0.8$ in increments of 0.4. Each of the 150 cells of the resulting 6 (sample size) x 5 (correlation) x 5 (distribution) design contains 5000 observations. Due to the large number of observations for each cell, we focus on effect size estimates (measured in terms of partial η^2) instead of *p*-values. In the analyses of the one-sample procedures, the source of variation explains virtually nothing of the variation in means of the simulated *t* and *z* distributions (R^2 measures < 0.001). The estimated partial η^2 measures also suggest that mean *t* and *z* values remain unaffected by the simulation parameters. The ANOVA results for the two-sample tests suggest that the strength of correlation has the largest impact on the mean *t* and *z* values. R^2 estimates for the two-sample *t* and *U* test were 0.141 and 0.143, respectively. The remaining factors of the simulations did not affect the distributions of the test statistics.

[3] See tables and figures at the end of this contribution.

Power

Tables 4 and 5 show the probabilities of rejecting the null hypothesis for a true effect of $\mu_1 - \mu_2 = 0.5$ (representing the power of the significance tests) for $n = 30$, 50, and 70. Again, results for the experimental conditions not shown are quite similar to the presented findings. Results suggest that, first, the power of all significance tests increases with the sample size and the true mean difference, as expected. Second, for normal and uniform deviates, the parametric procedures are generally more powerful than their nonparametric counterparts.

This pattern reverses for asymmetric distributions (Gumbel and exponential; see Table 5). In these cases, the *U* test is more powerful than the two-sample *t* test and the Wilcoxon test shows a power advantage over the paired *t* test. Furthermore, for the case of independent samples ($\rho = 0$) the two-sample *t* test is slightly more powerful than the paired *t* test. The same holds for the nonparametric procedures. Here, the *U* test consistently outperforms the Wilcoxon test. Comparing the power entries across the range of correlations implemented in the simulations, it becomes evident that the paired *t* test dramatically loses power in case of negatively correlated samples. The magnitude of the power loss varies with the magnitude of the negative correlation and holds for all considered distributions. The power loss is less pronounced for large true differences in means and large sample sizes. However, these are cases for which Schuster and Bortz (2010) recommend using the Wilcoxon test. However, the inspection of the power functions of the Wilcoxon test reveals a very similar power loss. Figures 2 – 4 show comparisons of the power curves of the Wilcoxon test and the paired *t* test for the four non-normal distributions and $\mu_1 - \mu_2 = 0.25$, 0.50, and 0.75, respectively. Each line represents the difference between the observed power curves based on 50,000 repetitions. Values above zero indicate a power advantage of the Wilcoxon test, values below zero indicate a power advantage of the paired *t* test. Apparently, the power superiority of the Wilcoxon tests depends on type of distribution, degree of correlation, sample size, and true underlying effect size. For uniformly distributed populations, the well-known power advantage of the paired *t* test is more pronounced for $\rho < 0$. For the logistic distribution, the Wilcoxon test is slightly more powerful than the paired *t* test. However, this power advantage decreases for strong positive correlations. For asymmetric distributions, the power differences tend to follow an inversely U-shaped curve. This implies that the advantage of the Wilcoxon test is more pronounced for moderately correlated samples. Most important is that the power advantage of the Wilcoxon test diminishes for negatively correlated samples.

Finally, to further explore the sensitivity of the simulated test statistics, ANOVAs were performed. Table 6 shows the ANOVA results for the power simulation again using the *t* and *z* values as dependent variables. ANOVAs were restricted to a 6 (sample size: $n = 30, \ldots(40)\ldots, 230$) x 5 (correlation: $\rho = -0.8, \ldots(0.4)\ldots, 0.8$) x 5 (distribution shapes) x 3 (effect size: $\mu_1 - \mu_2 = 0.25, 0.50, 0.75$) design. Again, each of the 450 cells contained 5000 observations.

Model fit estimates varied from $R^2 = 0.79$ to 0.90 depending on the significance test analyzed (see Table 6). Again, due to the large number of observations for each cell, we focus on partial η^2 estimates instead of *p*-values. As expected, the strongest effects result for sample size (partial η^2 values > 0.5), effect size ($\mu_1 - \mu_2$; all partial η^2 values > 0.7) and, only for the one-sample procedures, the correlation factor (partial η^2 values > 0.7). Average *t* and *z* values increase with the correlation between the two variables, the sample size, and the mean differences. In addition, the mean difference significantly interacts with the sample size factor in all ANOVA models (η^2 values range between 0.19 and 0.23). The analyses of the test statistics of the paired *t* test and Wilcoxon test further reveal meaningful 'effect × correlation' and 'sample size × correlation' interactions (partial η^2 values range between 0.10 and 0.38 and between 0.21 and 0.24, respectively; see Table 6). Figure 5 shows the average *t* and *z* values for the interaction effects. Apparently, test statistics of the paired *t* test and the Wilcoxon test increase with sample size and mean difference. This effect is even more pronounced for high positive correlations.

Discussion

Numerous previous studies dealt with robustness and power properties of parametric and nonparametric tests under non-normality, heterogeneity of variances, or a combination of both. Comparatively less attention has been paid to the behavior of the significance tests in cases where independence assumptions are violated (cf. von Eye, 2004). In particular, negative correlations seem underresearched. For a recently published exception see Zimmerman (2012). However, those parts of Zimmerman's study which considered negative correlations were restricted to normally distributed populations and sample sizes of 20, 25, 100, and 400. Furthermore, the study only focused on parametric significance tests. Using a more complex simulation design, the current study aimed at a more systematic evaluation of parametric and nonparametric tests developed for the one- as well as the two-sample problem under various degrees of negative and positive correlations, distribution shapes, sample sizes, and true differences in means.

Several conclusions can be drawn from the present results: First, as expected, those significance tests originally developed for the two-sample problem produce seriously biased decisions concerning the null hypothesis when samples are correlated. These results replicate earlier findings (e.g., Zimmerman et al., 1993). For example, even the small correlation of $\rho = -0.1$ between normally distributed samples produced Type I error rates which were outside the chosen robustness interval of 4 – 6%. This holds for both, the two-sample *t* test and the nonparametric *U* test, and implies that even modest departures from independence can make results of these tests hard to interpret. In addition, whether these tests are too liberal or too stringent depends on the sign of the population correlation, which is also in accordance with previous studies (Zimmerman, 1997, 2012).

In contrast, the significance tests developed for the one-sample problem kept the nominal significance level α over the entire range of the simulated correlations. We conclude that both procedures, the paired *t* test and the Wilcoxon test, have nonparametric properties with respect to between-group correlation. However, it is important to note that the tests still rely on the assumption of within-group independence. The nonparametric property of the paired *t* test does *not* hold for within-group correlations. Von Eye (1983, 2004) found that the Type I error rates of the one-sample *t* test are far above (below) the nominal significance level in cases of positive (negative) within-group correlations. Quite similar results were observed for the Wilcoxon test (Chlaß & Krüger, 2007).

Second, if samples are completely independent ($\rho = 0$) and $\mu_1 - \mu_2 > 0$ the two-sample *t* test constantly outperforms the paired *t* test because the two-sample *t* test employs $2n - 2$ degrees of freedom instead of $n - 1$ degrees of freedom of the paired *t* test. A quite similar effect can be observed for the *U* test and the Wilcoxon test which can be explained by the different ranking approaches and resulting number of possible ranks. The Wilcoxon test computes *n* signed ranks of difference scores. In contrast, the *U* test assigns 1, …, 2*n* possible ranks to the original scores, which increases the sensitivity for true differences (Iman, Hora & Conover, 1984). The power superiority of the *U* test might be better understood from the perspective of rank transformation theory (Conover & Iman, 1981). It can be shown that various nonparametric tests are asymptotically equivalent to the corresponding parametric tests performed on ranks replacing the original scores. Thus, the Wilcoxon test and the paired *t* test, performed on signed ranks replacing difference scores, will suggest equivalent statistical decisions. Similarly, the two sample *t* test, performed on ranks, only rejects the null hypothesis when the *U* test does. Hence, the explanation based on differences in degrees of freedom holds for the nonparametric procedures as well.

Third, the ANOVA results for the power simulation suggest that the power of the paired *t* test heavily depends on the true differences in means and the strength of correlation which is in line with previous results (e.g., Zimmerman, 1997). The latter can be explained by the fact that the variance of differences is defined as $\sigma_d^2 = \sigma_1^2 + \sigma_2^2 - 2 \cdot cov(x_1, x_2)$, where σ_1^2 and σ_2^2 denote the variances of x_1 and x_2, and $cov(x_1, x_2)$ denotes the covariance term of x_1 and x_2 (Hays, 1994). Thus, a strong positive relationship (i.e., a rather large covariance term) reduces the standard error of differences. In contrast, high negative correlations lead to rather large standard errors, which in turn lead to a loss in power to detect true differences. In these cases, Bortz and Schuster (2010) argue that the Wilcoxon test should be applied instead. The current results take exception to this recommendation. The simulated power functions of the Wilcoxon test show that the procedure also loses power when between correlations are negative. The inspection of Figures 2 – 4 shows that the power loss can sometimes be even greater than that of the paired *t* test. As a consequence, for uniformly distributed samples, the power advantage of the paired *t* test is even more pronounced than for negatively correlated samples. Furthermore, the well-known power superiority of the Wilcoxon test for skewed densities (cf. Blair & Higgin, 1985) diminishes for strong negative correlations in smaller samples. Therefore, researchers should not use the direction of correlation as a basis for choosing between the two procedures.

Finally, we hope that the insights from the current study might help to further improve already excellent introductory textbooks to statistics such as the one by Schuster and Bortz (2010). In addition, we would like to encourage future research on a family of significance tests developed by Zimmerman (2005, 2012), which currently seems to receive less attention in the social sciences. The basic idea behind these tests is to apply a modified version of the two-sample *t* test with a corrective term to account for the observed correlation between samples. The use of the corrective term seems to successfully resolve the problem of distorted Type I error rates. In addition, this modified two-sample *t* test shows a power advantage over the paired *t* test under several correlation scenarios due to the larger number of degrees of freedom. For a discussion concerning the larger number of degrees of freedom see also Wiedermann & Alexandrowicz (2011).

Acknowledgment

The authors are indebted to Rainer Alexandrowicz and Ingrid Koller for helpful discussions of earlier versions of this article.

Figures and tables

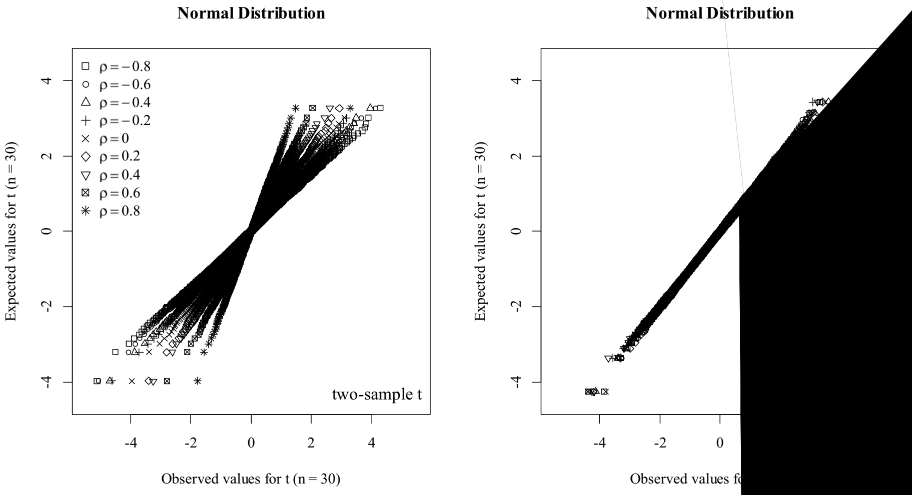

Figure 1:
Probability plots for the simulated *t* distributions as a function of populati

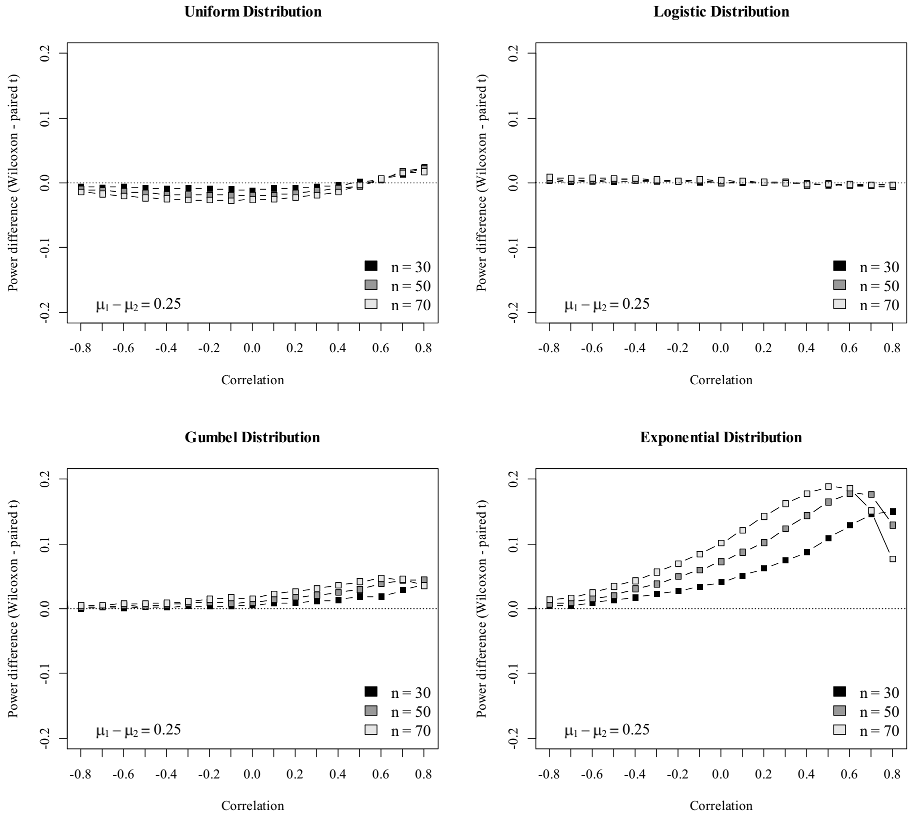

Figure 2:
Differences in empirical power values between the Wilcoxon test and the paired *t* test as a function of population correlation for non-normal distributions ($\mu_1 - \mu_2 = 0.25$).

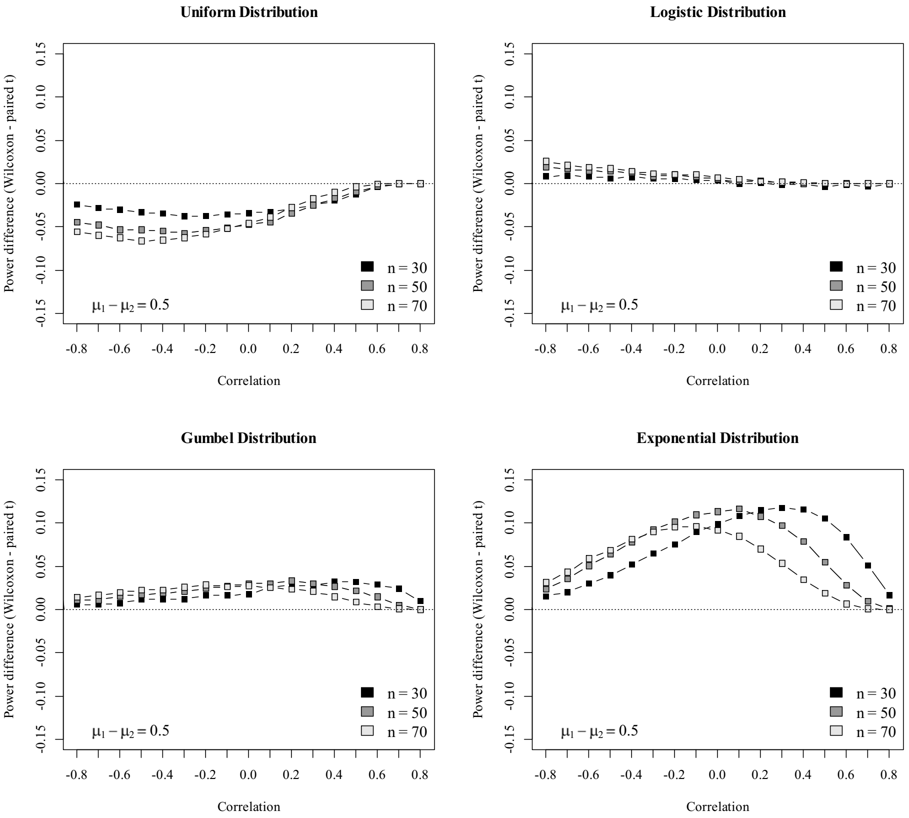

Figure 3:
Differences in empirical power values between the Wilcoxon test and the paired *t* test as a function of population correlation for non-normal distributions ($\mu_1 - \mu_2 = 0.50$).

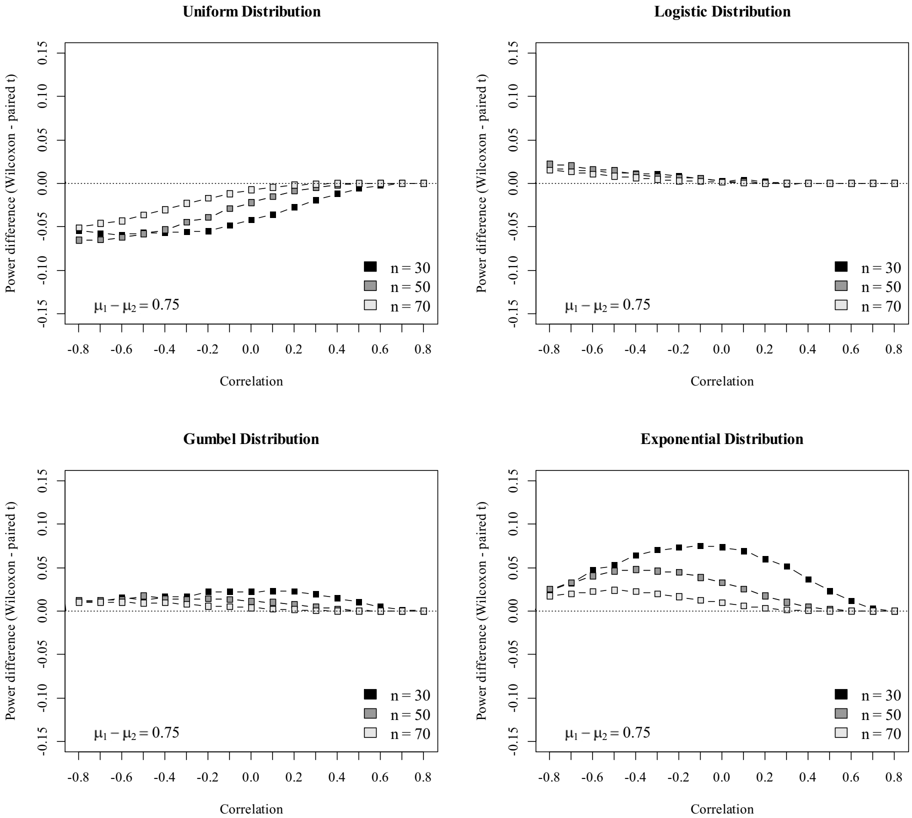

Figure 4:
Differences in empirical power values between the Wilcoxon test and the paired *t* test as a function of population correlation for non-normal distributions ($\mu_1 - \mu_2 = 0.75$).

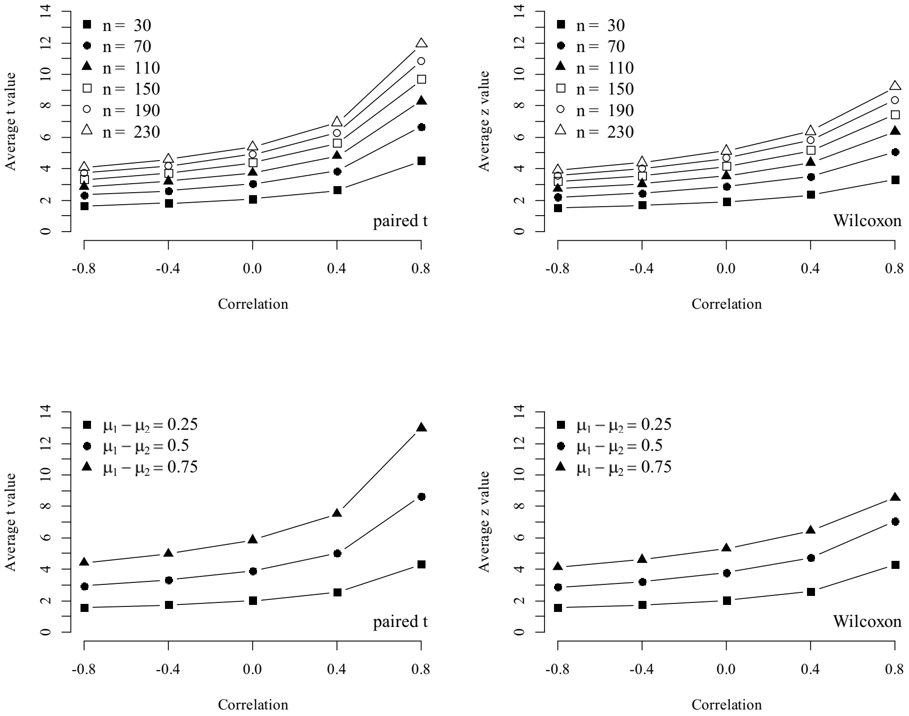

Figure 5:

Average test statistics of the paired *t* test and the Wilcoxon test as a function of correlation, sample size, and differences in means.

Table 1:
Type I error rates for symmetric population shapes (ρ = correlation, *t* = two-sample *t* test, paired *t* = paired *t* test, *U* = Wilcoxon-Mann-Whitney *U* test, *W* = Wilcoxon-matched-pairs-signed-ranks test, *n* = sample size).

	n = 30				*n* = 150				*n* = 250			
ρ	*t*	paired *t*	*U*	*W*	*t*	paired *t*	*U*	*W*	*t*	paired *t*	*U*	*W*
	Normal Distribution											
–0.8	0.142	0.049	0.141	0.048	0.146	0.047	0.142	0.050	0.148	0.054	0.146	0.054
–0.6	0.118	0.048	0.113	0.049	0.122	0.048	0.125	0.047	0.124	0.050	0.118	0.045
–0.4	0.097	0.048	0.091	0.047	0.103	0.050	0.100	0.050	0.096	0.048	0.094	0.050
–0.2	0.076	0.053	0.074	0.054	0.072	0.048	0.071	0.052	0.078	0.052	0.078	0.050
0	0.046	0.048	0.046	0.047	0.050	0.052	0.049	0.052	0.049	0.049	0.052	0.050
0.2	0.029	0.049	0.030	0.049	0.032	0.051	0.033	0.052	0.029	0.049	0.027	0.049
0.4	0.014	0.051	0.015	0.047	0.010	0.052	0.012	0.052	0.009	0.045	0.009	0.046
0.6	0.004	0.046	0.005	0.046	0.003	0.051	0.003	0.053	0.002	0.055	0.003	0.056
0.8	0.000	0.054	0.000	0.055	0.000	0.051	0.000	0.050	0.000	0.052	0.000	0.053
	Uniform Distribution											
–0.8	0.147	0.057	0.146	0.053	0.143	0.050	0.143	0.050	0.153	0.054	0.153	0.054
–0.6	0.118	0.048	0.117	0.047	0.113	0.050	0.114	0.049	0.114	0.046	0.114	0.047
–0.4	0.095	0.053	0.091	0.050	0.093	0.045	0.093	0.044	0.094	0.052	0.096	0.054
–0.2	0.069	0.048	0.068	0.051	0.068	0.049	0.068	0.048	0.071	0.048	0.072	0.049
0	0.049	0.047	0.047	0.048	0.054	0.054	0.052	0.054	0.052	0.053	0.054	0.051
0.2	0.031	0.053	0.029	0.054	0.029	0.051	0.028	0.049	0.024	0.047	0.024	0.048
0.4	0.012	0.048	0.014	0.047	0.013	0.051	0.012	0.054	0.011	0.051	0.012	0.052
0.6	0.004	0.053	0.005	0.051	0.003	0.049	0.003	0.048	0.003	0.050	0.003	0.051
0.8	0.000	0.044	0.000	0.045	0.000	0.050	0.000	0.054	0.000	0.052	0.000	0.055
	Logistic Distribution											
–0.8	0.141	0.046	0.138	0.047	0.140	0.048	0.139	0.045	0.143	0.047	0.138	0.046
–0.6	0.118	0.046	0.117	0.049	0.123	0.051	0.122	0.051	0.125	0.051	0.119	0.054
–0.4	0.095	0.048	0.093	0.049	0.103	0.049	0.094	0.049	0.097	0.051	0.090	0.048
–0.2	0.072	0.049	0.072	0.049	0.070	0.047	0.065	0.047	0.074	0.050	0.073	0.052
0	0.054	0.054	0.055	0.054	0.050	0.051	0.050	0.049	0.048	0.049	0.049	0.052
0.2	0.032	0.052	0.031	0.051	0.028	0.050	0.028	0.048	0.032	0.054	0.031	0.052
0.4	0.013	0.047	0.013	0.047	0.011	0.047	0.012	0.049	0.012	0.048	0.014	0.051
0.6	0.003	0.049	0.004	0.050	0.002	0.052	0.003	0.050	0.002	0.049	0.003	0.051
0.8	0.000	0.048	0.000	0.049	0.000	0.048	0.000	0.049	0.000	0.051	0.000	0.052

Table 2:
Type I error rates for asymmetric population shapes (ρ = correlation, *t* = two-sample *t* test, paired *t* = paired *t* test, *U* = Wilcoxon-Mann-Whitney *U* test, *W* = Wilcoxon-matched-pairs-signed-ranks test, *n* = sample size).

	n = 30				*n* = 150				*n* = 250			
ρ	*t*	paired *t*	*U*	*W*	*t*	paired *t*	*U*	*W*	*t*	paired *t*	*U*	*W*
	Gumbel Distribution											
–0.8	0.142	0.054	0.146	0.053	0.134	0.048	0.138	0.051	0.131	0.050	0.134	0.052
–0.6	0.120	0.047	0.121	0.047	0.118	0.048	0.122	0.046	0.121	0.054	0.120	0.056
–0.4	0.092	0.046	0.089	0.049	0.099	0.052	0.098	0.053	0.093	0.053	0.094	0.052
–0.2	0.067	0.047	0.067	0.044	0.071	0.049	0.074	0.047	0.069	0.048	0.069	0.047
0	0.049	0.050	0.050	0.051	0.052	0.052	0.050	0.049	0.050	0.049	0.053	0.045
0.2	0.029	0.048	0.031	0.048	0.034	0.060	0.033	0.059	0.026	0.049	0.030	0.046
0.4	0.015	0.049	0.015	0.050	0.014	0.050	0.015	0.054	0.013	0.050	0.013	0.050
0.6	0.004	0.048	0.006	0.051	0.002	0.047	0.003	0.051	0.001	0.047	0.002	0.049
0.8	0.000	0.047	0.001	0.049	0.000	0.049	0.000	0.049	0.000	0.049	0.000	0.047
	Exponential Distribution											
–0.8	0.123	0.045	0.157	0.044	0.120	0.043	0.161	0.042	0.126	0.047	0.163	0.049
–0.6	0.107	0.046	0.125	0.046	0.117	0.050	0.142	0.049	0.113	0.050	0.135	0.052
–0.4	0.093	0.049	0.101	0.052	0.091	0.048	0.104	0.049	0.093	0.050	0.111	0.052
–0.2	0.072	0.048	0.082	0.055	0.068	0.048	0.075	0.050	0.077	0.051	0.075	0.050
0	0.050	0.050	0.051	0.053	0.052	0.053	0.053	0.054	0.052	0.051	0.056	0.051
0.2	0.029	0.050	0.027	0.054	0.029	0.051	0.031	0.051	0.026	0.047	0.027	0.049
0.4	0.011	0.042	0.010	0.042	0.011	0.050	0.010	0.048	0.012	0.047	0.008	0.047
0.6	0.002	0.049	0.003	0.051	0.002	0.052	0.002	0.051	0.002	0.048	0.001	0.050
0.8	0.000	0.042	0.000	0.048	0.000	0.049	0.000	0.050	0.000	0.055	0.000	0.056

Table 3:
ANOVA results for the Type I error simulation.

Source	df	Type III Sum of Squares	Mean Squares	F-value	p-value	Partial η^2
two-sample *t* test ($R^2 = 0.140831$)						
Distribution	4	13.01	3.25	8.85	<.0001	0.000
Sample Size	5	16.71	3.34	9.10	<.0001	0.000
Distribution × Sample Size	20	7.07	0.35	0.96	0.507	0.000
Correlation	4	45051.67	11262.92	30654.50	<.0001	0.141
Distribution × Correlation	16	36.26	2.27	6.17	<.0001	0.000
Sample Size × Correlation	20	7.65	0.38	1.04	0.408	0.000
Distribution × Sample Size × Correlation	80	27.34	0.34	0.93	0.655	0.000
paired *t* test ($R^2 = 0.000299$)						
Distribution	4	0.80	0.20	0.53	0.712	0.000
Sample Size	5	34.53	6.91	18.4	<.0001	0.000
Distribution × Sample Size	20	6.81	0.34	0.91	0.577	0.000
Correlation	4	3.99	1.00	2.66	0.031	0.000
Distribution × Correlation	16	6.83	0.43	1.14	0.312	0.000
Sample Size × Correlation	20	6.71	0.34	0.89	0.595	0.000
Distribution × Sample Size × Correlation	80	24.54	0.31	0.82	0.881	0.000
Wilcoxon-Mann-Whitney *U* test ($R^2 = 0.143200$)						
Distribution	4	10.40	2.60	7.12	<.0001	0.000
Sample Size	5	1.56	0.31	0.85	0.512	0.000
Distribution × Sample Size	20	6.90	0.34	0.94	0.529	0.000
Correlation	4	45535.11	11383.78	31190.1	<.0001	0.143
Distribution × Correlation	16	148.02	9.25	25.35	<.0001	0.001
Sample Size × Correlation	20	9.90	0.50	1.36	0.132	0.000
Distribution × Sample Size × Correlation	80	29.36	0.37	1.01	0.465	0.000
Wilcoxon-matched-pairs-signed-ranks test ($R^2 = 0.000196$)						
Distribution	4	1.17	0.29	0.81	0.5165	0.000
Sample Size	5	1.88	0.38	1.05	0.389	0.000
Distribution × Sample Size	20	7.05	0.35	0.98	0.486	0.000
Correlation	4	3.45	0.86	2.39	0.0485	0.000
Distribution × Correlation	16	5.62	0.35	0.97	0.4818	0.000
Sample Size × Correlation	20	7.49	0.37	1.04	0.411	0.000
Distribution × Sample Size × Correlation	80	26.36	0.33	0.91	0.695	0.000

Table 4:
Relative frequencies of rejecting the null hypothesis for symmetric distributions (ρ = correlation, *t* = two-sample *t* test, paired *t* = paired *t* test, *U* = Wilcoxon-Mann-Whitney *U* test, *W* = Wilcoxon-matched-pairs-signed-ranks test, *n* = sample size, true differences in means: $\mu_1 - \mu_2 = 0.50$).

	$n = 30$				$n = 50$				$n = 70$			
ρ	*t*	paired *t*	*U*	*W*	*t*	paired *t*	*U*	*W*	*t*	paired *t*	*U*	*W*
	Normal Distribution											
–0.8	0.490	0.297	0.480	0.286	0.650	0.451	0.632	0.433	0.782	0.591	0.763	0.570
–0.6	0.487	0.320	0.474	0.306	0.661	0.497	0.641	0.477	0.782	0.633	0.767	0.614
–0.4	0.477	0.353	0.458	0.344	0.665	0.544	0.646	0.515	0.796	0.696	0.775	0.672
–0.2	0.476	0.394	0.455	0.382	0.689	0.620	0.669	0.596	0.812	0.753	0.792	0.733
0	0.482	0.468	0.459	0.453	0.702	0.695	0.680	0.678	0.833	0.831	0.822	0.812
0.2	0.475	0.555	0.456	0.538	0.704	0.775	0.683	0.757	0.862	0.908	0.839	0.890
0.4	0.462	0.675	0.447	0.654	0.750	0.890	0.715	0.869	0.894	0.966	0.874	0.954
0.6	0.462	0.842	0.436	0.828	0.784	0.973	0.752	0.968	0.938	0.997	0.915	0.996
0.8	0.459	0.988	0.448	0.984	0.855	1.000	0.830	1.000	0.983	1.000	0.972	1.000
	Uniform Distribution											
–0.8	0.488	0.284	0.462	0.259	0.656	0.457	0.623	0.414	0.767	0.582	0.729	0.532
–0.6	0.485	0.312	0.457	0.281	0.650	0.484	0.613	0.432	0.783	0.636	0.745	0.572
–0.4	0.474	0.339	0.436	0.304	0.663	0.543	0.626	0.482	0.790	0.686	0.747	0.616
–0.2	0.480	0.397	0.449	0.361	0.674	0.610	0.629	0.553	0.818	0.761	0.776	0.708
0	0.472	0.460	0.441	0.424	0.696	0.690	0.649	0.639	0.841	0.837	0.799	0.795
0.2	0.476	0.554	0.439	0.525	0.710	0.771	0.656	0.733	0.864	0.906	0.821	0.879
0.4	0.472	0.659	0.436	0.643	0.744	0.884	0.693	0.863	0.892	0.960	0.848	0.950
0.6	0.453	0.825	0.417	0.818	0.778	0.961	0.718	0.960	0.934	0.995	0.899	0.993
0.8	0.439	0.979	0.408	0.979	0.872	1.000	0.813	0.999	0.982	1.000	0.960	1.000
	Logistic Distribution											
–0.8	0.495	0.296	0.508	0.308	0.655	0.455	0.683	0.483	0.765	0.575	0.796	0.609
–0.6	0.490	0.317	0.501	0.330	0.659	0.501	0.687	0.517	0.787	0.637	0.813	0.655
–0.4	0.476	0.352	0.496	0.362	0.668	0.548	0.696	0.558	0.796	0.701	0.822	0.715
–0.2	0.487	0.409	0.509	0.413	0.686	0.619	0.718	0.624	0.814	0.756	0.848	0.769
0	0.468	0.450	0.492	0.458	0.693	0.684	0.730	0.692	0.834	0.829	0.862	0.834
0.2	0.480	0.558	0.510	0.558	0.716	0.784	0.750	0.790	0.854	0.896	0.882	0.898
0.4	0.481	0.686	0.512	0.681	0.749	0.889	0.783	0.888	0.897	0.966	0.928	0.966
0.6	0.483	0.831	0.517	0.831	0.782	0.971	0.819	0.972	0.933	0.995	0.950	0.996
0.8	0.468	0.985	0.525	0.984	0.849	1.000	0.894	1.000	0.978	1.000	0.987	1.000

Table 5:
Relative frequencies of rejecting the null hypothesis for asymmetric distributions (ρ = correlation, *t* = two-sample *t* test, paired *t* = paired *t* test, *U* = Wilcoxon-Mann-Whitney *U* test, *W* = Wilcoxon-matched-pairs-signed-ranks test, *n* = sample size, true differences in means: $\mu_1 - \mu_2 = 0.50$).

	n = 30				*n* = 50				*n* = 70			
ρ	*t*	paired *t*	*U*	*W*	*t*	paired *t*	*U*	*W*	*t*	paired *t*	*U*	*W*
	Gumbel Distribution											
–0.8	0.488	0.301	0.544	0.305	0.654	0.465	0.712	0.477	0.773	0.603	0.827	0.617
–0.6	0.480	0.327	0.539	0.338	0.658	0.514	0.720	0.521	0.787	0.656	0.849	0.681
–0.4	0.484	0.362	0.541	0.372	0.673	0.562	0.737	0.578	0.795	0.696	0.856	0.723
–0.2	0.490	0.415	0.552	0.432	0.677	0.609	0.753	0.639	0.818	0.769	0.885	0.797
0	0.475	0.464	0.536	0.486	0.699	0.692	0.780	0.727	0.836	0.834	0.898	0.858
0.2	0.493	0.558	0.562	0.586	0.718	0.777	0.802	0.808	0.851	0.893	0.912	0.914
0.4	0.491	0.673	0.571	0.703	0.739	0.870	0.828	0.899	0.889	0.956	0.945	0.972
0.6	0.499	0.833	0.586	0.859	0.780	0.962	0.870	0.976	0.924	0.994	0.973	0.997
0.8	0.477	0.974	0.619	0.984	0.842	0.998	0.930	1.000	0.969	1.000	0.996	1.000
	Exponential Distribution											
–0.8	0.498	0.319	0.708	0.333	0.661	0.487	0.873	0.511	0.786	0.629	0.948	0.662
–0.6	0.515	0.354	0.729	0.385	0.677	0.520	0.894	0.573	0.781	0.653	0.957	0.710
–0.4	0.513	0.377	0.737	0.434	0.673	0.560	0.896	0.635	0.806	0.718	0.969	0.794
–0.2	0.508	0.431	0.755	0.511	0.680	0.615	0.912	0.719	0.816	0.766	0.975	0.855
0	0.505	0.490	0.765	0.592	0.707	0.701	0.934	0.816	0.831	0.829	0.984	0.917
0.2	0.507	0.575	0.799	0.700	0.728	0.784	0.956	0.896	0.852	0.894	0.990	0.967
0.4	0.517	0.706	0.824	0.817	0.747	0.882	0.966	0.959	0.896	0.961	0.998	0.993
0.6	0.518	0.842	0.860	0.925	0.791	0.962	0.986	0.993	0.915	0.993	0.999	0.999
0.8	0.514	0.978	0.891	0.995	0.835	0.999	0.996	1.000	0.965	1.000	1.000	1.000

Table 6:
ANOVA results for the power simulation.

Source	df	Type III Sum of Squares	Mean Squares	F-value	p-value	Partial η^2
two-sample *t* test ($R^2 = 0.794149$)						
Effect size ($\mu_1 - \mu_2$)	2	5525038.70	2762519.40	2689175.00	<.0001	0.705
Distribution	4	686.11	171.53	166.97	<.0001	0.000
Effect size × Distribution	8	172.70	21.59	21.01	<.0001	0.000
Sample Size	5	2843210.70	568642.14	553545.00	<.0001	0.552
Effect size × Sample Size	10	542483.68	54248.37	52808.10	<.0001	0.190
Distribution × Sample Size	20	113.18	5.66	5.51	<.0001	0.000
Effect size × Distribution × Sample Size	40	91.86	2.30	2.24	<.0001	0.000
Correlation	4	529.32	132.33	128.82	<.0001	0.000
Effect size × Correlation	8	946.50	118.31	115.17	<.0001	0.000
Distribution × Correlation	16	98.41	6.15	5.99	<.0001	0.000
Effect size × Distribution × Correlation	32	34.31	1.07	1.04	0.399	0.000
Sample Size × Correlation	20	732.36	36.62	35.65	<.0001	0.000
Effect size × Sample Size × Correlation	40	792.50	19.81	19.29	<.0001	0.000
Distribution × Sample Size × Correlation	80	85.87	1.07	1.04	0.370	0.000
Effect size × Distribution x Sample Size × Correlation	160	178.96	1.12	1.09	0.209	0.000
paired *t* test ($R^2 = 0.904640$)						
Effect size ($\mu_1 - \mu_2$)	2	8353495.10	4176747.60	3493279.00	<.0001	0.756
Distribution	4	7878.65	1969.66	1647.35	<.0001	0.003
Effect size × Distribution	8	1382.17	172.77	144.50	<.0001	0.001
Sample Size	5	4251222.10	850244.42	711113.00	<.0001	0.613
Effect size × Sample Size	10	783273.67	78327.37	65510.10	<.0001	0.23
Distribution × Sample Size	20	153.77	7.69	6.43	<.0001	0.000
Effect size × Distribution × Sample Size	40	81.27	2.03	1.70	0.0038	0.000
Correlation	4	9492792.60	2373198.10	1984856.00	<.0001	0.779
Effect size × Correlation	8	1648188.70	206023.59	172311.00	<.0001	0.380
Distribution × Correlation	16	9976.03	623.50	521.47	<.0001	0.004
Effect size × Distribution × Correlation	32	1559.99	48.75	40.77	<.0001	0.001
Sample Size × Correlation	20	841075.12	42053.76	35172.20	<.0001	0.238
Effect size × Sample Size × Correlation	40	123571.40	3089.29	2583.77	<.0001	0.044
Distribution × Sample Size × Correlation	80	805.43	10.07	8.42	<.0001	0.000
Effect size × Distribution × Sample Size × Correlation	160	355.88	2.22	1.86	<.0001	0.000

continued

Wilcoxon-Mann-Whitney *U* test ($R^2 = 0.834770$)						
Effect size ($\mu_1 - \mu_2$)	2	4990991.50	2495495.80	2987530.00	<.0001	0.727
Distribution	4	669730.57	167432.64	200445.00	<.0001	0.263
Effect size × Distribution	8	33336.88	4167.11	4988.73	<.0001	0.017
Sample Size	5	3221592.30	644318.47	771358.00	<.0001	0.632
Effect size × Sample Size	10	506313.80	50631.38	60614.30	<.0001	0.212
Distribution × Sample Size	20	66637.47	3331.87	3988.82	<.0001	0.034
Effect size × Distribution × Sample Size	40	2118.69	52.97	63.41	<.0001	0.001
Correlation	4	127.89	31.97	38.28	<.0001	0.000
Effect size × Correlation	8	951.08	118.88	142.32	<.0001	0.001
Distribution × Correlation	16	83.77	5.24	6.27	<.0001	0.000
Effect size × Distribution × Correlation	32	106.12	3.32	3.97	<.0001	0.000
Sample Size × Correlation	20	323.13	16.16	19.34	<.0001	0.000
Effect size × Sample Size × Correlation	40	718.03	17.95	21.49	<.0001	0.000
Distribution × Sample Size × Correlation	80	100.49	1.26	1.50	0.0024	0.000
Effect size × Distribution × Sample Size × Correlation	160	167.05	1.04	1.25	0.0175	0.000
Wilcoxon-matched-pairs-signed-ranks test ($R^2 = 0.894070$)						
Effect size ($\mu_1 - \mu_2$)	2	4322918.60	2161459.30	3151347.00	<.0001	0.737
Distribution	4	84826.75	21206.69	30918.80	<.0001	0.052
Effect size × Distribution	8	2223.44	277.93	405.21	<.0001	0.001
Sample Size	5	3443666.80	688733.35	1004154.00	<.0001	0.691
Effect size × Sample Size	10	441486.86	44148.69	64367.60	<.0001	0.223
Distribution × Sample Size	20	8093.37	404.67	590.00	<.0001	0.005
Effect size × Distribution × Sample Size	40	246.06	6.15	8.97	<.0001	0.000
Correlation	4	4117025.40	1029256.40	1500627.00	<.0001	0.727
Effect size × Correlation	8	173295.84	21661.98	31582.60	<.0001	0.101
Distribution × Correlation	16	8224.22	514.01	749.42	<.0001	0.005
Effect size × Distribution × Correlation	32	11107.27	347.10	506.07	<.0001	0.007
Sample Size × Correlation	20	397857.90	19892.90	29003.30	<.0001	0.205
Effect size × Sample Size × Correlation	40	9938.24	248.46	362.24	<.0001	0.006
Distribution × Sample Size × Correlation	80	706.56	8.83	12.88	<.0001	0.001
Effect size × Distribution × Sample Size × Correlation	160	1027.32	6.42	9.36	<.0001	0.001

References

Bartlett, M. S. (1935). The effect of non-normality on the t-distribution. *Proceedings of the Cambridge Philosophical Society, 31*, 223-231.

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

Boneau, C. A. (1960). The effects of violation of assumptions underlying the t-test. *Psychological Bulletin, 57*, 49-64.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler (7. Aufl.)*. Berlin Heidelberg: Springer.

Chlaß, N., & Krüger, J. J. (2007). Small sample properties of the Wilcoxon signed rank test with discontinuous and dependent observations, *Jena Economic Research Papers*, No. 2007,032, http://hdl.handle.net/10419/25598

Cochran, W. G. (1947). Some consequences when the assumption for the analysis of variance are not satisfied. *Biometrics, 3*, 22-38.

Conover, W. J., & Iman, R. L. (1981). Rank transformation as a bridge between parametric and nonparametric statistics. *The American Statistician, 35*, 124-129.

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

Fisher, R. A. (1925). Applications of "Student's" distribution. *Metron, 5*, 90-104.

Gosset ("Student"), W. S. (1908). The probable error of a mean. *Biometrika, 6*, 1-25.

Guiard, V., & Rasch, D. (2004). The robustness of two sample tests for means: A reply on von Eye's comment. *Psychology Science, 46*, 549-554.

Hays, W. L. (1994). *Statistics (5th ed.)*. Wadsworth: Thompson Learning.

Herrendörfer, G., Rasch, D., & Feige, K. D. (1983). Robustness of statistical methods II. Methods of the one-sample problem. *Biometrical Journal, 25*, 327-343.

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

Hsu, P. L. (1938). Contributions to the theory of Student's t-test as applied to the problem of two samples. *Statistical Research Memoirs, 2*, 1-24.

Iman, R. L., Hora S. C., & Conover W. J. (1984). Comparison of asymptotically distribution-free procedures for the analysis of complete blocks. *Journal of the American Statistical Association, 79*, 674-685.

Lissitz, R. W., & Chardos, S. (1975). A study of the effect of the violations of the assumption of independent sampling upon the type one error rate of the two-sample t-test. *Educational and Psychological Measurement, 35*, 353-359.

Marsaglia, G., & Tsang, W. W. (2000). The Ziggurat method for generating random variables. *Journal of Statistical Software, 5*, http://www.jstatsoft.org/v05/i08/paper.

Moder, K. (2010). Alternatives to F-test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Testing and Assessment Modeling, 52*, 343-353.

Neave, H. R., & Granger, C. W. J. (1968). A monte carlo study comparing various two sample tests for differences in means. *Technometrics, 10*, 509-522.

Posten, H. O. (1978). The robustness of the two sample t-test over the Pearson system. *Journal of Statistical Computation and Simulation, 6*, 295-311.

Posten, H. O. (1979). The robustness of the one-sample t-test over the Pearson system. *Journal of Statistical Computation and Simulation, 6*, 133-149.

Posten, H. O. (1984). Robustness of the two-sample t-test. In D. Rasch & M. L. Tiku (eds.), *Robustness of statistical methods and nonparametric statistics* (p. 92-99). Dordrecht: D. Reidel Publishing Company.

Posten, H. O., Yeh, H. C., & Owen, D. B. (1982). Robustness of the two-sample t-test under violations of the homogeneity of variance assumptions. *Communications in Statistics: Theory and Methods, 11*, 109-126.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. New York: Wiley.

Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science, 46*, 175-208.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical Papers, 52*, 219-231.

Rasch, D., Kubinger, K. D., & Yanagida, T. (2011). *Statistics in Psychology using R and SPSS*. Chichester: Wiley.

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

Tuchscherer, A. & Pierer, H. (1985). Simulationsuntersuchungen zur Robustheit verschiedener Verfahren zum Mittelwertsvergleich im Zweistichprobenproblem (Simulationsergebnisse). [Simulation studies on robustness of several methods for the comparison of means in the two sample problem]. In P. E. Rudolph (ed.), *Robustheit V – Arbeitsmaterial zum Forschungsthema Robustheit. Probleme der angewandten Statistik*, 15, 1-42, Dummersdorf-Rostock.

von Eye, A. (1983). t-tests for single means of autocorrelated data – a simulation study. *Biometrical Journal, 25*, 801-805.

von Eye, A. (2004). Robustness is parameter-specific: A comment on Rasch and Guiard's robustness study. *Psychology Science, 46*, 544-548.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29*, 350-362.

Welch, B. L. (1947). The generalisation of Student's problem when several different population variances are involved. *Biometrika, 34*, 28-35.

Wiedermann, W., & Alexandrowicz, R. (2007). A plea for more general tests than those for location only: Further considerations on Rasch & Guiard's 'The robustness of parametric statistical methods'. *Psychology Science, 49*, 2-12.

Wiedermann, W., & Alexandrowicz, R. (2011). A modified normal scores test for paired data. *Methodology*, *7*, 25-38.

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

Zimmerman, D. W. (2005). Increasing power in paired-samples designs by correcting the Student t statistic for correlation. *Interstat,* http://interstat.statjournals.net/YEAR/2005/abstracts/0509002.php

Zimmerman, D. W. (2012). Correcting two sample z and t tests for correlation: An alternative to one sample tests on difference scores. *Psicológica*, *33*, 391-418.

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