A Comparison of Two-Sample Tests of Significance When Used With Variable Treatment Effects

Thomas W. O’GORMAN

This article compares the performance of many two-sample tests of significance that might be used to test the equality of means when the effect of the treatment is variable. Of the 19 tests that were compared, the normal scores test is recommended for general use in testing the null hypothesis of no treatment effect against the alternative that the distributions are stochastically ordered when the ratio of the larger standard deviation to the smaller standard deviation does not exceed 1.3. The Baumgartner-Weiß-Schindler tests and an adaptive test also have higher power than the pooled $t$-test, the unequal variance $t$-test, and the rank-sum test for many distributions. In the simulation studies, data in the first sample are generated from nine distributions, including long-tailed and skewed distributions. Data in the second sample are generated by adding a random treatment effect to a random variable that was generated from the same distribution that was used in the first sample. Because we restricted our power studies to treatment effects that are positive or zero, the population distributions will be stochastically ordered. The results of these studies demonstrate that the normal scores test is often more powerful than the $t$-tests and the rank-sum test. If the ratio of the standard deviations does exceed 1.3, then one of the $t$-tests is recommended.

Key Words: Normal scores test; Permutation test; Wilcoxon test.

1. Introduction

In this research, we will compare 19 two-sample tests to determine which one can be recommended in experimental settings when the treatment effect may be variable. In many experiments, subjects have some variability in their response to treatments. For example, in an experiment designed to measure the effect of exercise on blood pressure, Georgiades et al. (2000) reported differences in variability between the control group and the exercise-diet group with blood pressure measurements that were recorded during mental stress. In our investigation of the significance level and power of these tests, we will not assume that treatments are constant, which would imply a location shift model.

There are several reasons why the location shift model may not be realistic. The experimental units may have responses that are monotone functions of follow-up time, but the functions are not identical so for a fixed follow-up time the responses will vary. Another source of variability in response in human and animal experiments could be differences in baseline characteristics (e.g., initial weight). In this article, we will investigate the relative power of many tests of significance when the location shift model may be violated.

Before we describe the test procedures that we will compare, we need to clearly describe the null and
alternative hypotheses. In an experimental setting, the null hypothesis is that the treatment has no effect, so the distributions in the first and second populations are identical. Let \( X_1, \ldots, X_m \) be a random sample from the first population that has a cumulative distribution function (cdf) of \( F(x) \), and let \( Y_1, \ldots, Y_n \) be a random sample from the second population that has cdf \( G(y) \). A random variable \( X \) with cdf \( F(x) \) is said to be stochastically smaller than the random variable \( Y \) with cdf \( G(y) \) if \( F(x) \geq G(x) \), with \( F(x) > G(x) \) for at least one \( x \)-value. If the random variable \( X \) is stochastically smaller than the random variable \( Y \), then these distributions are said to be stochastically ordered. See Randles and Wolfe (1979, p. 130) for details.

If the treatment effect is equal to the same positive constant for all subjects then \( F(x) \) and \( G(y) \) are stochastically ordered, with \( G(y) \) shifted to the right of \( F(x) \). If the treatment effect is not constant, but is positive for some subjects and is negative for others, then the distributions will be stochastically ordered. In some studies it may be reasonable to assume that the treatment effect increases with time and that the variability of the treatment effect increases with time. Because the researcher would typically not know the direction of the effect, we will assume, at the time the measurements are recorded, that the treatment effect is nonnegative for all subjects or nonpositive for all subjects. Thus, we will perform a two-sided test of

\[ H_0 : F(x) = G(y) \text{ for every } x \]

versus

\[ H_a : F(\cdot) \text{ is stochastically smaller, or larger, than } G(\cdot). \]

Clearly, stochastically ordered distributions would result from positive treatment effects in an experimental setting, but it is also possible to perform the same tests with observational data if the distributions are stochastically ordered. While the stochastically ordered alternative is more general than the shift alternative, it is more restrictive than the general alternative \( H_a : F(x) \neq G(x) \) for at least one \( x \). In this research we focus on the performance of two-sample tests with stochastically ordered distributions, but it should be remembered that many of these tests were designed for a shift alternative and that some of the other tests were designed to be powerful with a general alternative. The Behrens-Fisher problem, which assumes unequal variances under the null hypothesis, will not be addressed because, in an experimental setting, the null hypothesis usually is a statement that the treatment has no effect, which implies that \( F(x) = G(x) \) for all \( x \).

2. Description of Two-Sample Tests

The pooled (or equal variance) \( t \)-test is one of the most widely used tests of significance because it is the most powerful test when both populations are normal and the variances are equal. In addition, simulation studies have shown that this test maintains its level of significance with many nonnormal distributions when the variances are equal. The problem with the pooled \( t \)-test is that it is not as powerful as the Wilcoxon rank-sum test (RS test) and the normal scores test (NS test) when the distributions are skewed. In addition, it is not known how the pooled \( t \)-test performs with variable treatment effects.

The unequal variance \( t \)-test is recommended by some researchers because it does not assume that the variances are equal. However, O’Gorman (1997) demonstrated that this test does not maintain its level of significance with skewed distributions when the sample sizes are not equal. In an effort to avoid this difficulty, we have included a permutation version of this test. This permutation Welch \( t \)-test is performed by computing the Welch test statistic in the usual manner, but the \( p \)-value is estimated by taking many random permutations of the group labels, and for each of these permutations the Welch test statistic is computed. These test statistics are used to compute a permutation distribution. Typically, the values of \( m \) and \( n \) are so large that a full permutation distribution would be difficult to obtain, so we usually use a sample permutation method with a large number of permutations. Let \( R \) be the number of permutations and let \( E \) be the number of permutation test statistics that are as extreme or more extreme than the unpermuted test statistic. Following Davidson and Hinkley (1997, p. 140), the two-sided \( p \)-value \((p)\) will be estimated by \( p = (E + 1)/(R + 1) \). This test will be called the permutation Welch \( t \)-test. In this study, we used \( R = 2000 \) for all tests that require a permutation approach to obtain a \( p \)-value. We also used this approach to construct a permutation version of the pooled \( t \)-test, which will be called the permutation \( t \)-test. We used \( R = 2000 \) for the sample permutation tests because it would give a reasonably accurate estimate of the \( p \)-value that would be obtained from a full permutation distribution. This value of \( R \) will give a standard error of 0.005 for \( p \)-values near \( \alpha = 0.05 \). In practice, when only one dataset is analyzed, a larger value of \( R \) could be used.

Some researchers prefer to use nonparametric tests to perform two-sample tests because, by relying on the ranks, the test statistics are not too sensitive to outliers. We included the Wilcoxon rank-sum test that uses the ranks as scores and we used the large sample normal approximation to obtain the two-sided \( p \)-value. These procedures are described by Randles and Wolfe (1979, chap. 9) and by Hájek, Sidák, and Sen (1999, chap. 4). We have also included the Van der Waerden version of the normal score test. If we let \( R_i \) be the rank in the combined sample for the \( i \)-th observation, the normal scores are given by \( \Phi^{-1}[R_i/(m + n + 1)] \), where \( \Phi(\cdot) \) is the cdf of the standard normal distribution. These normal scores are used
A Comparison of Two-Sample Tests of Significance When Used With Variable Treatment Effects

along with a large sample normal approximation to obtain the two-sided \( p \)-value. It is also possible to use normal scores and rank scores in a Mantel-Haenszel procedure to perform these rank tests. While the Mantel-Haenszel tests are more commonly used with categorical data, they can be used to perform the rank test if the ranks are used as column scores and the sample indicator is used as the row scores. The choice of column scores is discussed as column scores and the sample indicator is used as the

can be used to perform the rank test if the ranks are used

to perform these rank tests. While the Mantel-Haenszel

tests are more commonly used with categorical data, they

are variable, the population distributions will differ, and

it is reasonable to include test statistics that are sensitive
to differences between the cumulative distribution functions. Let \( F_m(x) \) and \( G_n(y) \) denote the empirical cdf’s for the \( m \) samples from the first population and the \( n \) samples from the second population, respectively. The two-sample version of the Cramér-von Mises test statistic, which was studied by Rosenblatt (1952), is

\[
CVM_2 = \frac{mn}{(m+n)^2} \left\{ \sum_{i=1}^{m} [F_m(x_i) - G_n(x_i)]^2 + \sum_{j=1}^{n} [F_m(y_j) - G_n(y_j)]^2 \right\}.
\]

The exact distribution of \( CVM_2 \) has been tabulated for small samples, but we have used a permutation test to estimate the \( p \)-value. A variant of the \( CVM_2 \) test that was investigated by Schmid and Trede (1995) has a test statistic

\[
CVM_1 = \frac{(mn)^{1/2}}{(m+n)^{3/2}} \left\{ \sum_{i=1}^{m} |F_m(x_i) - G_n(x_i)| + \sum_{j=1}^{n} |F_m(y_j) - G_n(y_j)| \right\}.
\]

We use a permutation method to obtain the \( p \)-value for this test. It was expected that the \( CVM_2 \) test and the \( CVM_1 \) test would be sensitive to location and scale, so they might have high power to detect a variable treatment effect.

Several tests have been proposed that are adaptive, which means that the procedures used to perform the tests depend on the empirical distribution of the combined samples. The oldest and best known of these is the HFR test that was proposed by Hogg, Fisher, and Randles (1975). This test maintains its level of significance because all of the component tests are rank tests and the selection of the rank scores depends on selection statistics that are functions of the order statistics of the combined samples, which are independent of the ranks for every distribution. For details see Randles and Wolfe (1979, chap. 11). In a small simulation study, this test has been shown by Hogg, Fisher, and Randles (1975) to have higher power than some other tests.

Another adaptive test was proposed by O’Gorman (2012, chap. 3). In this test, weights are assigned to individual observations, with outliers getting smaller weights, and a weighted least squares approach (WLS) is used to compute an \( F \)-test statistic based on the weighted observations. To ensure that the test maintains its significance level, a permutation method is used to compute the \( p \)-value.

Several two-sample tests for changes in location and scale have been proposed. Lepage (1971) proposed a test that combines the test statistic of the Wilcoxon rank-sum
Table 1. Distributions used to generate individual effects

| Distribution            | Description                                                                 |
|-------------------------|-----------------------------------------------------------------------------|
| Uniform                 | Uniform on $[-\sqrt{3}, \sqrt{3}]$                                          |
| Normal                  | Standard normal                                                             |
| $t$ with df = 4         | $t$ with df = 4, adjusted to give $\sigma = 1$                              |
| Symmetric bimodal       | $0.5 N(-\sqrt{9}/3, 4/13) + 0.5 N(\sqrt{9}/3, 4/13)$                       |
| Skewed low kurtosis     | RST distribution with $\alpha_1 = 1$ and $\alpha_2 = 4.2$                  |
| Skewed high kurtosis    | RST distribution with $\alpha_1 = 1$ and $\alpha_2 = 8.4$                  |
| Highly skewed low kurtosis | RST distribution with $\alpha_1 = 2$ and $\alpha_2 = 11.4$              |
| Highly skewed high kurtosis | RST distribution with $\alpha_1 = 2$ and $\alpha_2 = 15.6$            |
| Skewed bimodal          | $0.75 N(-1/2, 1/7) + 0.25 N(3/2, 4/7)$                                      |

NOTE: All distributions have a mean of zero and standard deviation of one. The bimodal distributions are mixtures of normals. For RST distributions, see Ramberg et al. (1979) for details.

3. Description of Simulation

To evaluate the relative power of these tests of significance, we used nine distributions to generate the data in the first sample. These individual effect distributions are listed in Table 1 along with a brief description of each. All nine of these distributions had $\mu = 0$ and $\sigma = 1$. In the table, we use $N(\mu, \sigma^2)$ to denote the normal distribution with mean $\mu$ and variance $\sigma^2$. For the $t$ distribution, we generated a random variable from the $t$ distribution with df = 4; then it was divided by $\sqrt{2}$ so that it had unit variance. The observations from the symmetric bimodal distribution are generated by a $N(-\sqrt{9}/3, 4/13)$ distribution with probability one-half, and from a $N(\sqrt{9}/3, 4/13)$ distribution otherwise. Most of the skewed distributions are members of the Ramberg, Schmeiser, and Tukey (RST) family of distributions, which allows the user to specify the first four moments. For these distributions we set $\mu = 0$, $\sigma^2 = 1$, and we specified the skewness ($\alpha_1$) and kurtosis ($\alpha_4$) in the manner described by Ramberg et al. (1979). The observations from the bimodal skewed distribution were generated by a $N(-1/2, 1/7)$ distribution with probability 0.75 and from a $N(3/2, 4/7)$ distribution otherwise.

While the observations in the first sample were generated from one of the individual effect distributions in Table 1, the observations in the second sample are obtained as the sum of the individual effect and the treatment effect. The individual effect is obtained from one of the nine distributions listed in Table 1 and the treatment effect is generated from a treatment distribution and is independent of the individual effect. We used five treatment effects for each sample size configuration and distribution. Let $t_i$ be the treatment effect for the $i$th observation and let $\Delta$ be the expected value of the treatment effect. If we let $\epsilon_i$ be the individual effect for the $i$th observation, then the observations in the first sample are generated from $y_i = \mu + t_i + \epsilon_i$ for $i = 1, \ldots, m$, and the observations in the second sample are generated from $y_i = \mu + t_i + \epsilon_i$ for $i = 1, \ldots, n$. For the constant treatment effect we set
For the second treatment effect, the effect is uniformly distributed so that \( t_i \sim U[0, 2\Delta] \) for \( i = 1, \ldots, n \). For the third treatment effect, the effect is distributed as an exponential with mean \( \Delta \) so that \( t_i \sim \exp(\Delta) \) for \( i = 1, \ldots, n \). To generate even more variability in the second sample, the treatment effects for the fourth effect are generated from \( U[0, 4\Delta] \) with probability one-half, and are set to zero with probability one-half.

For the last treatment effect, the effects are generated from \( \exp(2\Delta) \) with probability one-half, and are set to zero with probability one-half.

If \( \Delta \) is set to a large value then most of the tests will have high power, so it will be difficult to compare the powers of the tests. For most of the simulations we investigated the relative powers of these tests when \( \Delta \) was set to obtain empirical powers near 80%. In these studies we use \( \alpha = 0.05 \), we set the desired power (\( \Pi \)) such that \( \Pi = 0.8 \), and we determine

\[
\Delta = \left[ \Phi^{-1}(1 - \alpha/2) - \Phi^{-1}(1 - \Pi) \right] \sqrt{1/m + 1/n} \tag{1}
\]

to approximate the desired power when a constant treatment effect is used. For given sample sizes of \( m \) and \( n \), and a desired power of \( \Pi \), we compute \( \Delta \) and use it with the five treatment effects. We do not adjust \( \Delta \) to compensate for the loss of power that occurs with variable treatment effects nor do we adjust it to compensate for nonnormal data.

When variable treatment effects are used, the second population will have greater variability than the first population. Table 2 gives the population standard deviations in the second group for several combinations of sample size and treatment effects for a desired power of 80%. Because \( \Delta \) will decrease as the sample size increases, the variability of the treatment effect will decrease with increasing sample size. Because the treatment effects and the individual effects are independent, the standard deviations of the observations in the second population were obtained as the square root of the sum of the variance of the individual effect and the variance of the treatment effect.

We used total sample sizes of \( N = m + n = 30, 60, 120, \) and 240 in these simulations. For each \( N \), we used a balanced size of \( m = n = N/2 \). We also used unbalanced sizes of \( (m = N/3, n = 2N/3) \) and \( (m = N/5, n = 4N/5) \), so that the larger sample would have the larger variability. To compare the tests, when the smaller sample had the larger variability, we used \( (m = 2N/3, n = N/3) \) and \( (m = 4N/5, n = N/5) \). With four levels of \( N \) and five sample size configurations for each \( N \), we have investigated \( 4 \times 5 = 20 \) sample size configurations. Because we used nine error distributions for each sample size configuration and used five treatment effects for each distribution, we have performed \( 4 \times 5 \times 9 \times 5 = 900 \) simulation studies for the power comparisons.

### 4. Results

We begin by describing the level of significance of these tests. In Table 3 the empirical level of significance is shown, based on simulations that use 10,000 datasets, for two sample size configurations with a normal distribution, a skewed low kurtosis distribution, and a skewed high kurtosis distribution. For each of these tests, we determined if the \( p \)-value was less than \( \alpha \) for each of the 10,000 datasets, and then we determined the proportion of datasets that led to rejection of the null hypothesis, which is the empirical level of significance. The majority of tests have empirical significance levels near the \( \alpha = 0.05 \) level, while a few have levels that are below the \( \alpha = 0.05 \) level. Note that the six tests for location and scale generally have empirical significance levels below the \( \alpha = 0.05 \) level for \( m = n = 15 \). However, the Welch \( t \)-test had an empirical significance level of 5.68% when \( m = 12 \) and \( n = 48 \) observations were generated from a skewed, low kurtosis distribution. The online supplementary tables give the empirical significance level for \( N = 60 \) for all distributions and these results show that the empirical significance level of the Welch \( t \)-test equals 6.86% for the bimodal skewed distribution when \( m = 12 \) and \( n = 48 \). We conclude that the Welch \( t \)-test does not always maintain its significance level. Note that the results from simulation studies for the significance level with \( m = 12 \) and \( n = 48 \) also apply to studies having \( m = 48 \) and \( n = 12 \) because they are performed under the null hypothesis.

Because most of the tests maintain the nominal level of significance, we will proceed to display, in Table 4, the empirical powers of the 19 tests that were used in six of these simulations. We used 10,000 datasets to calculate the empirical power in these simulations. In Table 4, the empirical powers are shown for a simulation study with a constant treatment effect with \( m = n = 15 \) observations that were generated from a normal distribution. Because the treatment effect was constant, these are the powers for testing the null hypothesis against a shift alternative. We used Equation (1) to determine the value of \( \Delta \) that would
Table 3. Empirical level of significance in percent for a normal, a skewed low kurtosis (Low K.), and a skewed high kurtosis (High K.) individual effect distribution

| Name of test            | Distribution | m = n = 15 | m = 12, n = 48 |
|-------------------------|--------------|------------|----------------|
|                        |              | Normal     | Skewed low K.  | Normal     | Skewed low K.  | Skewed high K. |
| Pooled t                |              | 4.91       | 5.03           | 4.64       | 5.00           | 4.87           | 4.86           |
| Permutation t           |              | 5.02       | 5.15           | 5.05       | 5.06           | 5.05           | 4.87           |
| Welch t                 |              | 4.79       | 4.80           | 4.54       | 5.27           | 5.68           | 5.03           |
| Permutation Welch t     |              | 4.98       | 5.04           | 5.02       | 5.22           | 4.95           | 5.11           |
| Wilcoxon rank-sum (RS)  |              | 4.56       | 5.23           | 5.25       | 4.74           | 4.44           | 4.66           |
| Normal scores (NS)      |              | 4.57       | 4.87           | 5.01       | 5.06           | 4.86           | 4.93           |
| Baumgartner (BWS)       |              | 4.48       | 5.16           | 5.12       | 4.98           | 4.56           | 4.86           |
| Fligner-Policello (FP)  |              | 4.57       | 5.18           | 5.28       | 4.95           | 4.57           | 5.10           |
| CVM1                    |              | 4.84       | 5.50           | 5.34       | 5.05           | 4.66           | 4.92           |
| HFR adaptive            |              | 4.49       | 4.93           | 5.15       | 4.73           | 4.69           | 4.56           |
| Adaptive WLS            |              | 4.75       | 4.99           | 5.12       | 5.12           | 4.89           | 4.98           |
| Wilcoxon-AB             |              | 4.14       | 4.52           | 4.73       | 4.57           | 4.56           | 4.67           |
| Wilcoxon-Mood           |              | 3.94       | 4.43           | 4.61       | 4.51           | 4.49           | 4.62           |
| NS-Klotz                |              | 3.31       | 3.62           | 3.93       | 4.22           | 4.02           | 4.40           |
| Quadratic raw data      |              | 3.57       | 3.81           | 3.41       | 4.75           | 4.92           | 5.08           |
| Quadratic rank          |              | 4.39       | 5.01           | 5.26       | 4.76           | 4.85           | 4.91           |
| Quadratic NS            |              | 3.96       | 4.19           | 4.50       | 4.42           | 4.31           | 4.63           |
| Neuhäuser               |              | 4.81       | 5.21           | 5.08       | 5.24           | 4.46           | 4.96           |

achieve 80% power for the $t$-test with constant treatment effects. If $m = n = 15$ with data generated from a normal distribution we find that, as expected, the pooled $t$-test was the most powerful test against a shift alternative. However, many of the other tests had powers within 5% of the maximum. Note that the last seven tests in Table 4 had lower power than the other tests. The next two columns give the same information for datasets that were generated from skewed distributions having $\alpha_3 = 1$. With these skewed distributions, the empirical powers of these tests show that the nonparametric tests tend to be more powerful than the $t$-tests. In the last three columns, we see that these general results also hold for larger samples having $m = n = 30$ when the treatment effect is variable. In general, the $t$-tests have greater power with normally distributed errors and the nonparametric tests have greater power with skewed errors. Because, for $m = n = 30$, the treatment effect distribution was $U[0, 2\Delta]$, the standard deviation in the second population is 1.084, as was shown in Table 2. Note that the results for $m = n = 15$ are not comparable to those for $m = n = 30$ because different values of $\Delta$ were used for the simulations.

For each of the four sample sizes with five sample configurations of balanced and unbalanced data, we obtained the population standard deviation in the second group, which we will denote by $\sigma$. In our initial analysis of the results, we found that for simulations that had $\sigma \leq 1.3$ the results were similar to those that had $\sigma = 1.0$. We will give some results for the simulations having $\sigma > 1.3$ at the end of this section, but most of the simulations had $\sigma \leq 1.3$, as we observed in Table 2. For example, with $N = 60$, the constant, $U[0, 2\Delta]$, and Exponential($\Delta$) treatments effects had $\sigma \leq 1.3$, and with $N = 240$ all of the treatment effects had $\sigma \leq 1.3$. Because it would be impractical to show all of the results from this study, in Table 5 we give the average power over the five sample configurations and the nine error distributions for simulation studies that have $\sigma \leq 1.30$. The results for all simulations having $N = 60$ are given in the online supplementary tables.

To make the results more manageable, we will eliminate from further consideration those tests that have lower power than the other tests, and those tests that appear to have very similar performance to other well-known tests. In Tables 4 and 5, we observe that the last seven tests have lower power than many of the traditional tests. This includes the Wilcoxon-AB test, the Wilcoxon-Mood test, the NS-Klotz test, the three quadratic tests, and the Neuhäuser test. Because these seven tests have lower power than many of the other tests, we will not consider them further. Nor does there seem to be a good reason to include, in our comparisons, the permutation version of the pooled $t$-test because the pooled $t$-test maintains its level of significance at or near $\alpha$ with a variety of distributions and sample size configurations. On the other hand, the Welch’s $t$-test does not maintain its significance level with unbalanced sample sizes and skewed data, as we observed in Table 3. Hence, we will use the permutation Welch $t$-test because it maintains its significance level, and we will not consider the Welch $t$-test further.
Table 4. Empirical power in percent for a normal, a skewed low kurtosis (Low K.), and a high kurtosis (High K.) individual effect distribution for a desired power near 80%

| Treatment effect | m = n = 15 | m = n = 30 |
|------------------|------------|------------|
| Constant         | Normal     | Skewed low K. | Skewed high K. | Normal     | Skewed low K. | Skewed high K. |
| Pooled t         | 77.3       | 76.6       | 78.2       | 74.9       | 75.7       | 76.4       |
| Permutation t     | 77.4       | 76.6       | 78.5       | 74.8       | 75.9       | 76.6       |
| Welch t           | 77.1       | 76.4       | 77.8       | 74.9       | 75.6       | 76.3       |
| Wilcoxon rank-sum (RS) | 75.4       | 80.7       | 85.1       | 74.8       | 75.9       | 83.7       |
| Normal scores (NS) | 75.7       | 80.9       | 82.7       | 72.6       | 79.9       | 81.1       |
| Baumgartner (BWS) | 73.2       | 82.8       | 84.8       | 73.9       | 79.9       | 83.0       |
| Fligner-Policello (FP) | 75.4       | 80.1       | 84.4       | 72.6       | 79.9       | 83.5       |
| CVM1             | 72.6       | 81.3       | 85.5       | 70.6       | 78.6       | 83.8       |
| CVM2             | 74.9       | 81.5       | 85.6       | 72.3       | 79.8       | 84.1       |
| HFR adaptive     | 73.9       | 80.9       | 83.6       | 72.5       | 79.1       | 82.2       |
| Adaptive WLS     | 76.4       | 82.7       | 83.5       | 73.8       | 81.1       | 81.2       |
| Wilcoxon-Mood    | 61.3       | 77.2       | 76.3       | 62.7       | 72.1       | 74.0       |
| Wilcoxon-Mood    | 60.4       | 77.1       | 76.3       | 62.8       | 71.7       | 74.1       |
| NS-Klotz         | 59.8       | 76.0       | 74.0       | 62.9       | 71.6       | 70.9       |
| Quadratic raw data | 62.5       | 77.9       | 75.0       | 64.0       | 72.2       | 70.7       |
| Quadratic rank   | 62.8       | 78.9       | 78.2       | 63.6       | 72.5       | 74.8       |
| Quadratic NS     | 62.3       | 78.0       | 75.9       | 64.0       | 72.3       | 71.9       |
| Neuhäuser        | 69.0       | 74.9       | 80.0       | 65.6       | 68.9       | 75.7       |

NOTE: The treatment effect parameter \( \Delta \) was determined to give approximately 80% power.

We note, in Table 5, that the average power of the HFR adaptive test is more than 3% below that of the adaptive WLS test for all sample sizes. In addition, the average power of the HFR test is less than the average power of the rank-sum test for all sample sizes. Hence, we will not include the HFR test in further comparisons.

We also want to exclude any tests whose performance is closely approximated by a standard test. To find tests whose performance is closely related to other tests, we computed the Pearson correlations between the powers of all tests over the 612 simulation studies having \( N = 30, 60, 120, \) and \( 240 \) with \( \sigma \leq 1.3 \). We found a correlation of \( r = 0.996 \) between the Wilcoxon rank-sum test and the CVM1 test and a correlation of \( r = 0.982 \) between the Wilcoxon rank-sum test and the CVM2 test. We also note that the average power of the rank-sum test approximates the average power of the CVM1 test and that the CVM2 test is less powerful than the Wilcoxon test. Hence, we will not include either of the Cramér-von Mises tests in the other comparisons. None of the other tests had correlations with the \( t \)-tests or the rank-sum test that exceeded \( r = 0.973 \). After excluding these tests, we have seven tests remaining that will be compared.

In Table 6 we see the average power, when \( \Delta \) is set to achieve 80% power, for the seven most promising tests for
simulations that have greatest power, over the seven tests, for the 27 balanced simulations. The results show that if a researcher had a variety of datasets similar to those used in this study, and if they choose to use the NS test to analyze every dataset, then they would be assured that they were using the most powerful test, or nearly the most powerful test, of these seven tests. In contrast, if they had decided to use one of the $t$-tests for all datasets having $m = n$, they would often be using a test that had much lower power than some of the other tests. Hence, for datasets having $N = 60$ and $m = n$ the pooled $t$-test and the Welch $t$-test appear to be poor choices for general use, and the NS, BWS, and adaptive tests appear to be good choices.

If $N = 60$ and $m < n$, the larger sample will have the larger population variance. Figure 2 gives the percent of the simulations where the tests have power within 5% of the greatest power for simulations designed to give approximately 80% power with $N = 60$ and $m < n$ with $\sigma \leq 1.3$. In these simulations, the BWS test that was proposed by Baumgartner, Weiß, and Schindler (1998) is superior to the other tests, but the NS test and the adaptive test also have high power. Again, the $t$-tests are not good choices when $m < n$.

Figure 1 gives the percent of the simulations where the tests have the greatest power, or within 5% of the greatest power, over the seven tests, for the 27 balanced simulations that have $\Delta$ set to give approximately 80% power with $N = 60$ and $m = n$ with $\sigma \leq 1.3$. These results show that if a researcher had a variety of datasets similar to those used in this study, and if they choose to use the NS test to analyze every dataset, then they would be assured that they were using the most powerful test, or nearly the most powerful test, of these seven tests. In contrast, if they had decided to use one of the $t$-tests for all datasets having $m = n$, they would often be using a test that had much lower power than some of the other tests. Hence, for datasets having $N = 60$ and $m = n$ the pooled $t$-test and the Welch $t$-test appear to be poor choices for general use, and the NS, BWS, and adaptive tests appear to be good choices.

If $N = 60$ and $m < n$, the larger sample will have the larger population variance. Figure 2 gives the percent of the simulations where the tests have power within 5% of the greatest power for simulations designed to give approximately 80% power with $N = 60$ and $m < n$ with $\sigma \leq 1.3$. In these simulations, the BWS test that was proposed by Baumgartner, Weiß, and Schindler (1998) is superior to the other tests, but the NS test and the adaptive test also have high power. Again, the $t$-tests are not good choices when $m < n$.

Figure 2. The percentage of simulations that had empirical power within 5% of the greatest power for the seven tests of significance. These percentages are computed from the 45 simulations that have $N = 60$ and $m < n$ with $\sigma \leq 1.3$. For these simulations the larger group had the greater variability.

Table 6. Average power in percent for tests over nine individual effect distributions and treatment effects having $\sigma \leq 1.3$ for total sample sizes $N = m + n$

| Test         | Pooled t | Perm. Welch Rank-Sum | NS | BWS | FP | Adaptive |
|--------------|----------|-----------------------|----|-----|----|----------|
| $N = 30 m = n$ | 74.7     | 74.9                  | 77.7 | 77.7 | 78.4 | 77.3  | 78.9   |
| $m < n$      | 72.6     | 71.1                  | 73.9 | 74.9 | 78.9 | 70.3  | 76.2   |
| $m > n$      | 74.5     | 67.6                  | 75.4 | 75.2 | 70.4 | 73.8  | 78.0   |
| $N = 60 m = n$ | 75.0     | 75.1                  | 78.8 | 80.2 | 80.3 | 78.8  | 80.9   |
| $m < n$      | 74.7     | 74.2                  | 78.2 | 80.0 | 82.2 | 76.4  | 80.7   |
| $m > n$      | 75.9     | 73.0                  | 79.4 | 80.4 | 77.6 | 78.9  | 82.1   |
| $N = 120 m = n$ | 76.0     | 76.0                  | 79.9 | 81.6 | 82.3 | 79.9  | 82.0   |
| $m < n$      | 73.8     | 75.2                  | 77.6 | 79.1 | 82.0 | 77.5  | 78.8   |
| $m > n$      | 75.9     | 72.3                  | 77.8 | 80.1 | 77.9 | 76.7  | 80.6   |
| $N = 240 m = n$ | 76.8     | 76.8                  | 81.0 | 84.0 | 85.4 | 81.0  | 85.2   |
| $m < n$      | 75.1     | 76.8                  | 79.3 | 81.6 | 84.3 | 79.8  | 82.1   |
| $m > n$      | 76.9     | 74.4                  | 79.1 | 83.0 | 82.7 | 78.0  | 84.1   |

NOTE: The treatment effect parameter $\Delta$ was determined to give approximately 80% power. In each row, the maximum average power and those within 3% of the maximum average power are given in bold.
If \( m > n \), the larger sample will have the smaller population variance. For simulations having \( N = 60 \) and \( m > n \) with \( \sigma < 3 \), that were designed to give approximately 80% power, the percent of the simulations where the tests have power within 5% of the greatest power are shown in Figure 3. The results show that the adaptive test is the best choice if the larger sample has the smaller population variance.

So far, we have limited the choice of \( \Delta \) to a value that would produce an empirical power near 80% for the pooled \( t \)-test. To see if these results will apply to other values of \( \Delta \), we performed simulations having \( N = 60 \) with desired powers near 50% and 90%. The average power for simulations having \( N = 60 \) with \( \sigma < 1.3 \) are given in Table 7. The average powers demonstrate that the NS test is more powerful, on average, than the pooled \( t \)-test, the Welch \( t \)-test, and the rank-sum test. It also shows that, when \( m < n \), the BWS test is the most powerful test and when \( m > n \) the NS test and the Adaptive tests are more powerful than the other tests. The relative power of the tests with 50% power and 90% power are similar to those for 80% power that are shown in Table 6.

We also wanted to determine which tests would be appropriate if the treatment effect was so variable that \( \sigma > 1.3 \). In Table 8, we show the average power over the simulations designed to give approximately 80% power that had \( \sigma > 1.3 \), which in this study only occurs with smaller samples that have highly variable treatment effects that are listed in Table 2. The results show that, if \( \sigma > 1.3 \), the \( t \)-tests tend to be more powerful than the other tests. When the larger sample has the greater variability the Welch \( t \)-test is more powerful and when the smaller sample has the greater variability the pooled \( t \)-test is more powerful. If the sample sizes are equal the Welch \( t \)-test is slightly more powerful, on average.

Because the results for \( \sigma > 1.3 \) differ greatly from those for \( \sigma < 1.3 \), we investigated the relationship between \( \sigma \) and the relative power of these tests. If \( m = n \)

---

**Table 7.** Average power in percent for tests over nine distributions with \( \sigma < 1.3 \) for total sample sizes \( N = m + n = 60 \)

| Test | Pooled Perm. Welch BWS NS FP Adaptive |
|------|--------------------------------------|
| Power = 50% | | |
| \( m = n \) | 46.6 | 46.8 | 51.2 | 52.1 | 52.4 | 51.2 | 53.3 |
| \( m < n \) | 43.6 | 46.8 | 48.4 | 49.5 | 53.2 | 47.4 | 49.7 |
| \( m > n \) | 47.1 | 40.3 | 49.2 | 50.4 | 46.0 | 48.4 | 52.5 |
| Power = 90% | | |
| \( m = n \) | 87.2 | 87.3 | 90.3 | 91.3 | 91.8 | 90.3 | 92.1 |
| \( m < n \) | 86.2 | 84.8 | 88.4 | 89.8 | 91.6 | 86.7 | 90.5 |
| \( m > n \) | 87.0 | 85.2 | 89.3 | 90.2 | 88.2 | 88.7 | 91.3 |

NOTE: The treatment effect parameter \( \Delta \) was determined to give approximately 50% power for some simulations and 90% power for other simulations. In each row, the maximum average power and those within 3% of the maximum average power are given in bold.

---

**Table 8.** Average power in percent for tests over nine distributions with highly variable treatment effects so that \( \sigma > 1.3 \)

| Test | Pooled t Perm. Welch RS NS BWS FP Adaptive |
|------|--------------------------------------|
| \( m = n \) | 56.0 | 56.8 | 49.7 | 51.9 | 54.7 | 49.6 | 49.2 |
| \( m < n \) | 42.3 | 58.8 | 44.9 | 44.5 | 48.4 | 48.0 | 40.4 |
| \( m > n \) | 60.9 | 42.1 | 46.5 | 51.7 | 50.5 | 41.6 | 49.2 |

NOTE: The treatment effect parameter \( \Delta \) was determined to give approximately 80% power. In each row, the maximum average power and those within 3% of the maximum average power are given in bold.

123
5. Conclusions

We have attempted to evaluate the power of many two-sample tests of significance when the treatment effect is variable. If the treatment effect is constant, or somewhat variable so that the ratio of standard deviations is less than 1.3, we find that:

- The NS test, the BWS test, and the adaptive test are often more powerful than the pooled t-test, the permutation Welch t-test, the Wilcoxon rank-sum test, and the FP test.
- The BWS test often has lower power than the NS test and the adaptive test if \( m > n \).
- There is little difference in power between the NS and adaptive tests.
- The power advantage of the NS test decreases with the ratio of the standard deviations.

Consequently, if the ratio of standard deviations does not exceed 1.3 then the NS test and the Adaptive test are recommended. However, the NS test has several advantages over the adaptive test. The NS test is described in many nonparametric textbooks and is relatively easy to understand. Not only does the NS test have high power over a range of distributions and treatment effects, but also it maintains its level of significance without using a permutation method. In addition, it is available in statistical software packages, including the NPAR1WAY procedure in SAS/STAT© (SAS Institute 2014) software. In contrast, the Adaptive test requires a permutation-testing method to obtain a \( p \)-value. The major advantage of the Adaptive test is that it can also be used for more complex designs, as described by O’Gorman (2012).

These results apply to datasets where it is reasonable to assume that the distributions are stochastically ordered. If a study is being planned, results from similar studies or pilot studies should be used, when available, to determine if it is reasonable to assume that the ratio of standard deviations will be 1.3 or less. If the data have already been obtained, an analysis strategy might be to first decide if it is reasonable to assume that the treatment effect, if any, is one-directional. If the treatment effect is nonnegative for all subjects, or nonpositive for all subjects, then the distributions should be stochastically ordered and our results are relevant. Let \( \hat{\omega} \) be the ratio of the largest sample standard deviation to the smallest sample standard deviation. If \( \hat{\omega} \leq 1.3 \), then the normal scores test would be a good choice. If \( \hat{\omega} > 1.3 \), then the treatment effect is highly variable so that one of the \( t \)-tests would be recommended.

In some research we assume that the treatment effect is constant so we can estimate the mean treatment effect, which will be the individual treatment effect. If we are...
unwilling to assume that the treatment effect is constant, we would not be able to estimate a treatment effect for an individual, but would be able to estimate the mean treatment effect.

If observational data are obtained from two populations, the results of this research may be of some value provided the distributions are stochastically ordered. However, if the distributions are not stochastically ordered, with either experimental or observational data, then these results are not relevant. If the distributions are not ordered, the analyst may decide that a test for location is not appropriate and may consider one of the tests for location and scale such as the test of Lepage (1971) or the test by Neuhäuser, Leuchs, and Ball (2011).

References

Baumgartner, W., Weiß, P., and Schindler, H. (1998), “A Nonparametric Test for the General Two-Sample Problem,” *Biometrics*, 54, 1129–1135. [117,122]

Davidson, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge, UK: Cambridge University Press. [116]

Durán, B. S., Tsai, W. S., and Lewis, T. O. (1976), “A Class of Location-Scale Nonparametric Tests,” *Biometrika*, 63, 173–176. [118]

Fligner, M. A., and Policello, G. E. (1981), “Robust Rank Procedures for the Behrens-Fisher Problem,” *Journal of the American Statistical Association*, 76, 162–168. [117]

Georgiades, A., Sherwood, A., Gullette, E. C. D., Babyak, M. A., Hindelliter, A., Waugh, R., Tweedy, D., Craighead, L., Bloomer, R., and Blumenthal, J. A. (2000), “Effects of Exercise and Weight Loss on Mental Stress-Induced Cardiovascular Responses in Individuals With High Blood Pressure,” *Hypertension*, 36, 171–176. [115]

Graubard, B. I., and Korn, E. L. (1987), “Choice of Column Scores for Testing Independence in Ordered 2 × K Contingency Tables,” *Biometrics*, 43, 471–476. [117]

Hájek, J., Sidák, Z., and Sen, P. K. (1999), *Theory of Rank Tests* (2nd ed.), London: Academic Press. [116]

Hogg, R. V., Fisher, D. M., and Randles, R. H. (1975), “A Two-Sample Adaptive Distribution-Free Test,” *Journal of the American Statistical Association*, 70, 656–661. [117]

Klotz, J. (1962), “Nonparametric Tests for Scale,” *Annals of Mathematical Statistics*, 34, 498–512. [118]

Lepage, Y. (1971), “A Combination of Wilcoxon’s and Ansari-Bradley’s Statistics,” *Biometrika*, 58, 213–217. [117,125]

Neuhäuser, M. (2003), “A Note on the Exact Test Based on the Baumgartner–Weiß–Schindler Statistic in the Presence of Ties,” *Computational Statistics and Data Analysis*, 42, 561–568. [117]

Neuhäuser, M., Leuchs, A., and Ball, D. (2011), “A New Location-Scale Test Based on a Combination of the Ideas of Levene and Lepage,” *Biometrical Journal*, 53, 525–534. [118,125]

O’Brien, P. C. (1988), “Comparing Two Samples: Extensions of the t, Rank-Sum, and Log-Rank Tests,” *Journal of the American Statistical Association*, 83, 52–61. [118]

O’Gorman, T. W. (1997), “A Comparison of an Adaptive Two-Sample Test to the t-test and the Rank-Sum Test,” *Communications in Statistics—Simulation and Computation*, 26, 1393–1411. [116]

—— (2012), *Adaptive Tests of Significance Using Permutations of Residuals With R and SAS*, New York: Wiley. [117,124]

Ramberg, J. S., Dudewicz, E. J., Tadikamalla, P., R., and Mykytka, E. F. (1979), “A Probability Distribution and Its Uses in Fitting Data,” *Technometrics*, 21, 201–214. [118]

Randles, R. H., and Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Statistics*, New York: Wiley. [116,117]

Rosenblatt, M. (1952), “Limit Theorems Associated With Variants of the von Mises Statistic,” *Annals of Mathematical Statistics*, 23, 617–623. [117]

SAS Institute (2014), *SAS/STAT User’s Guide, Version 13.2*, Cary, NC: SAS Institute. [124]

Schmid, F., and Trede, M. (1995), “A Distribution Free Test for the Two Sample Problem for General Alternatives,” *Computational Statistics and Data Analysis*, 20, 409–419. [117]

About the Author

Thomas W. O’Gorman is Associate Professor, Division of Statistics, Northern Illinois University, DeKalb, IL 60115 (E-mail: twogorman@gmail.com).