A MONTE CARLO SIMULATION STUDY OF THE PERFORMANCE OF HYPOTHESIS TESTS UNDER ASSUMPTION VIOLATIONS

GARETH CLEVELAND
Colby College

Abstract

Hypothesis testing is frequently utilized in a wide range of disciplines as researchers attempt to draw inferences from data. Although most hypothesis tests theoretically require certain assumptions for their accuracy, these assumptions are often not known or simply ignored. The robustness of a test is defined as the ability of the test to withstand assumption violations with respect to its Type I error probability. This study seeks to empirically investigate the robustness of several two-sample hypothesis tests. Through Monte Carlo simulations, exact Type I error rates are calculated for five different tests under a wide range of simulation setups. Two of the five tests are parametric: Student's t-test and Welch's t-test. The remaining three tests are distribution free: the Mann-Whitney U test, a bootstrap-based test, and a permutation test. Results demonstrate that Welch's t-test, the bootstrap test, and the permutation test perform reasonably well under a variety of assumption violations.

"I can't believe schools are still teaching kids about the null hypothesis. I remember reading a big study that conclusively disproved it *years* ago."
- xkcd comic

I. Introduction

Those who have taken a statistics course, as well as many who have not, have no doubt used a hypothesis test. It would seem that every introductory class on applied statistics spends considerable time on significance testing, giving the misconception that everything one learns can be ultimately summarized by a p-value. In fact, many are likely under the impression that without hypothesis testing, there would be no use for the discipline of statistics. After all, how can one draw meaningful conclusions without a formal rejection of a null hypothesis?

With modern technology, hypothesis testing, one type of statistical inference, has become accessible to a far broader population than just the well-versed statistician. The pandemic of big data is sweeping the globe, and many are struggling to keep up. Thus today hypothesis testing is readily available for the novice and expert alike, whether on a smart phone or one of many software packages.

So with endless amounts of data and easy access to software, conducting inferential procedures has become convenient. And what is convenient is bound to be abused. Professionals in many fields are able to run analyses on their data without being held accountable for quantitative or qualitative accuracy. Appreciation for the technicalities of practicing inferential statistics in general is lacking.

Hypothesis testing is a method for determining how well theory fits with observation or vice versa (Wackerly, Mendenhall, & Scheaffer, 2008). Of course, inference procedures are based fundamentally on probability. A hypothesis is tested by determining how likely it is to be true based on observation or sampling. Every hypothesis test has four essential components: a null hypothesis, an alternative hypothesis, a test statistic, and a rejection region (Wackerly et al., 2008). The null hypothesis, denoted H_0, is typically the default condition which researchers are seeking evidence against. The alternative hypothesis, denoted H_A, is what researchers are seeking evidence to support. Thus in general, inference is structured around whether there is enough evidence to suggest that H_A is more likely to be true then H_0. The test statistic is the measure used to determine how much evidence exists against H_0. It is calculated from the observed data and examined relative to the rejection region. The rejection region is the range of test statistics which are extreme enough to lead to a conclusion that there is reason to doubt H_0. If the test statistic falls into the rejection region, there is significant evidence to suggest that H_0 is not true. However, if a test statistic does not fall into the rejection region, the conclusion is not acceptance of H_0, but instead that there is a lack of evidence against it.

A hypothesis test may be either parametric or nonparametric in nature. Parametric tests are often concerned with the values of population parameters, such as the mean, median, or variance since they assume that the underlying distribution of the population can be modeled through the knowledge of one or more parameters. Therefore H_0 and H_A are often concerned with the value of a given parameter. Parametric procedures are applicable whenever "the distribution(s) from which the samples(s) is (are) taken is (are) specified except for the values of a finite number of parameters" (Wackerly et al., 2008). In other words, whenever the population parameters are at least partly known, the population distribution is at least partly specified, and parametric hypothesis testing may be appropriate.

Nonparametric testing involves more generalized null hypotheses. It may involve testing some quality of a sample that cannot be measured by a parameter or testing whether two distributions are the same. In contrast to parametric procedures, nonparametric procedures are applicable when sampling distributions are not well-specified. Not enough is known about the population to make inferences which rely on its parameters or distribution.

Inherent in a discussion about parametric and nonparametric testing is the recognition that testing hypotheses often requires certain assumptions to be met. This is always true in the parametric case, where knowledge about the population distribution allows for the test statistic to have certain properties. Knowing these properties allows the researcher to estimate the likelihood of obtaining a given statistic and whether or not that statistic falls into the rejection region, provided certain assumptions are made about distributional parameters. Nonparametric tests require fewer, if any, assumptions in order to be conducted. To test a general hypothesis about the shape of a distribution, it makes sense that the only assumption needed is that the distribution exists. It is not necessary to know if the distribution is normal or, when testing a single population, what its moments are equal to. Perhaps the primary value of nonparametric tests is that they are readily available when parametric tests cannot be utilized. When the assumptions that a parametric test rely on for its accuracy are violated, a nonparametric test may be more valid.

With any test, two types of errors can be made when deciding whether to reject H_0. A *Type I error* occurs when H_0 is rejected in favor of H_A, but H_0 is actually true. A *Type II error* occurs when H_0 is not rejected but should be because H_A is true. For a given test, Type I and Type II errors happen with certain probabilities, α and β, respectively. It is important to note that the researcher selects α based on the desired size of the rejection region. This is because α is also the significance level of a test, the maximum probability at which a test statistic is extreme enough to be rejected. In the parametric case, for a given α, distributional assumptions allow for determination of the rejection region prior to running the test. However, often with nonparametric tests the rejection region is unknown until after the test has been conducted. This is because a theoretical distribution of the test statistic is generated in the process of conducting the test. Although not examined in this study, it should also be mentioned that $1 - \beta$ is known as the power of the test, or the probability of rejecting a false H_0.

Of course the theoretical assumptions needed for hypothesis testing are frequently not met. For example, distributions are not quite shaped according to their models, samples are not quite randomly selected, or it is not known whether parameters actually possess the necessary properties. All tests are based on some statistical and probability theory, which guarantees that they work as advertised. This theory is dependent on the assumptions which are specified with each test. But what happens when circumstances dictate that assumptions for a test do not hold? Is that test still viable? If a test behaves well even without its assumptions being met, it is referred to as *robust*. Otherwise, it is *nonrobust*. Unsurprisingly, tests often fall somewhere in between, and can be more sensitive to some assumption violations than others. Thus, it is often necessary to refer to a test's *robustness*. Robustness is typically measured by the accuracy of the error probabilities above. For example, a researcher may intend to set a rejection region with $\alpha = 0.05$, but if the test is relatively nonrobust, the true rejection region may be much larger or much smaller. A test has *undercoverage* if the true rejection region, and equivalently the true α, is too large. If the true α is too small, *overcoverage* is present. Undercoverage implies excessive Type I error, whereas overcoverage implies excessive Type II error. Because Type I error is generally viewed as being more serious than Type II error, overcoverage is preferable to undercoverage in most circumstances.

This returns the discussion to the previously articulated convenience of hypothesis testing. Prior research has demonstrated that many common tests are in fact nonrobust to assumption violations. Thus inference is not appropriate in such circumstances and due to the presence of undercoverage or overcoverage, there is a large degree of uncertainty surrounding the validity of conclusions.

This paper seeks to empirically investigate the robustness of a handful of two-sample hypothesis tests. Through Monte Carlo simulations, exact Type I error rates are calculated for five different tests under a wide range of simulation setups. Two of the five tests are parametric: Student's t-test and Welch's t-test. The remaining three tests are nonparametric: the Mann-Whitney U test, a bootstrap-based test, and a permutation-based test. For each test, the same factors are manipulated to form different simulation setups: sample size, standard deviation, and skewness. The standard deviation of a distribution measures average variability in the population, while the skewness indicates the degree of asymmetry present. In every simulation, H_0 is true, which allows for determination of Type I error rates. Also, for each test $\alpha = 0.05$, which provides the standard against which simulation results can be compared.

II. The Tests

Two-Sample Student's t-test

The t-test is perhaps the most widely used and written about parametric test. Student's t, as it is often called, is useful when sample sizes are too small for the distribution of the test statistic to approximate a standard normal distribution. In the two-sample case, which we concern ourselves with, the statistic is,

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\overline{X}$ is the mean from the first sample and $\overline{Y}$ is the mean from the second sample. The sample sizes are denoted by n_1 in the first sample and n_2 in the second. The term S_p is the pooled estimator for the standard deviation, σ, of the two samples. It is given by,

$$S_p = \sqrt{\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}}$$

where S_X^2 and S_Y^2 are the sample variances for the first and second samples, respectively. If H_0 is true and the difference in the population means is zero, T follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom. For a given significance level, α, H_0 is rejected if T is greater than t_α or less than t_α for a one-sided H_A or if $|T| > t_{\alpha/2}$ for a two-sided H_A. Here, t_α (or $t_{\alpha/2}$) is the critical value that defines the rejection region for significance level α, which is the most extreme value which T can take and still not lead to rejection of H_0. (Wackerly et al., 2008).

There are three assumptions for the two-sample t-test. The samples should be drawn randomly so that observations are jointly independent. They should also be both drawn from normal population distributions. Thirdly, we must assume that these distributions have equal variances.

Whether the independence assumption is met can typically only be inferred from the study design and sampling methodology used. The reason for the assumption is that to obtain an accurate comparison of two samples (and the underlying populations) the samples must not be related. The difference in means between two samples cannot be estimated if one sample depends on the other sample because the estimate of the difference will depend on covariance between the populations.

The need for normally distributed populations results from the need for T to follow a t-distribution. If both populations are normally distributed, then the difference in means between the two populations is also normally distributed. This normally distributed difference in means is necessary for T to follow a t-distribution. This can be seen in an expansion of the definition of T:

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \left[\frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right] \Bigg/ \sqrt{\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2(n_1 + n_2 - 2)}} = \frac{Z}{\sqrt{\frac{W}{v}}}$$

Z has a standard normal distribution with standard deviation σ. W has a χ^2 distribution $v = n_1 + n_2 - 2$ degrees of freedom. Thus the statistic T is formed out of a standard normal and a χ^2 distribution (Wackerly et al., 2008).

The equal variance assumption is important because for T to follow a t-distribution, a single quantity must be specified for the variance of $\overline{X} - \overline{Y}$. Thus it is assumed that the variances of both samples are the same: $\sigma_X^2 = \sigma_Y^2$. However, we do not know the values of σ_X^2 and σ_Y^2, so we must estimate the common variance (σ^2) by pooling information from both samples. This is done through the calculation of S_p^2 above, which is simply the weighted average of the sample variances, S_X^2 and S_Y^2. The problem of testing a difference in population means in the presence of unknown population variances is known as the *Behren's Fisher Problem* (Cressie & Whitford, 1986). If $\sigma_X^2 \neq \sigma_Y^2$, S_p is not an accurate estimate of σ^2 and T may not follow a t-distribution. Thus the equal variance assumption is necessary for the t-test in its standard form.

There is an extensive literature on the robustness of the t-test to assumption violations, particularly the second and third of those listed above. Cressie and Whitford (1986) show that if normality is violated, adjustments can be made to critical values or significance levels so that the t-statistic can still be utilized. However, if sample sizes are unequal, robustness to the equal variance assumption is lost and an alternative statistic, Welch's T discussed below, should be used. The issue of unequal sample sizes leading to diminished robustness is also discussed in Posten, Yeh, & Owen (1982), which finds decreased performance of two-sample t-tests as sample sizes become more disparate. Furthermore, Ramsey (1980) concludes that even with normal populations, equal sample sizes, and lenient standards, "the t test cannot be considered completely robust to unequal variances." He goes on to provide guidelines for when the t-test is appropriate to use for normal populations with differing variances. Sawilowsky and Blair (1992) examine robustness to the normality assumption by running simulations under eight different distribution shapes. The authors find generally that robustness to Type I error holds for equal sample sizes, large samples, and two-tailed tests but often not for distributions with "extreme skewness." However, according to Lumley et al. (2002), the normality assumption is not necessary for large samples – less than 500 in the most extreme cases. It is of course more crucial for small samples.

Welch's t-test

To deal with unequal variances in the two samples – the Behrens-Fisher problem – a new, intended to be robust, statistic was proposed by Welch (1938). We will denote it T_W and it is defined as follows:

$$T_W = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}$$

where $\overline{X}$, $\overline{Y}$, S_X^2, S_Y^2, n_1, and n_2 are defined as before. T_W approximates a t distribution with e degrees of freedom if H_0 is true, where e is defined as,

$$e = \frac{\left(\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}\right)^2}{\left(\frac{S_X^4}{{n_1}^2(n_1 - 1)} + \frac{S_Y^4}{{n_2}^2(n_2 - 1)}\right)}$$

Here, $e$is always less than the $n_1 + n_2 - 2$ degrees of freedom associated with the distribution of T. This theoretically serves to offset the problems that arise when a common variance between the populations is not assumed. Although T_W does not follow a t-distribution with $n_1 + n_2 - 2$ degrees of freedom, it does approximate a t-distribution with $e$degrees of freedom (Cressie & Whitford, 1986).

Welch's t-test has the same independence and normality assumptions as Student's t-test. The difference of course is the lack of the homogeneity of variance assumption – which is by design. Welch's t-test does not require population variances to be equal.

Research has shown T_W to be considerably more robust than T to violations of the equal variance assumption. Ramsey (1980) suggests Welch's t-test is a viable alternative to the t-test when variances are unequal. The disadvantages are a loss of power if variances do happen to be equal and the fact that T_W merely approximates a t-distribution, but is not guaranteed to be exact. Cressie & Whitford (1986) recommend T_W in cases of unequal variances and a further adjustment to the statistic in situations where normality is also violated. They also advise against using T_W when variances are equal due to the diminished power mentioned above. Welch's T may often be overlooked and underutilized (Ruxton, 2008). The author advocates widespread use of the test due to its robustness to both the equal variance and normality assumptions and its advantages over nonparametric methods.

Mann-Whitney U Test

The t-tests above are used to test the equality of location parameters (i.e., the mean) under assumptions about variances and distributions. The Mann-Whitney U (MWU) test, also known as the Wilcoxon rank-sum test, provides a nonparametric alternative to the t-test. The test is designed to detect differences in the locations of two distributions without specifying what those distributions look like. It is a rank-based test, meaning calculations are done with the observations' ranks rather than with their original values.

To conduct a MWU test, the observations from both samples are first pooled and then ranked. If there are n total observations, the largest in magnitude is assigned a rank of n while the smallest is given a rank of 1. The statistic U is obtained as follows: for each observation from the second sample, the number of observations from the first sample which precede it in rank is counted. The statistic is the sum of the counts calculated for each second sample observation. If its value is very large or very small, it may be an indication that the population means are different. However, for large sample sizes, these manual calculations are tedious and instead it is preferable to use the formula for the U statistic:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - W$$

where n_1 is the first sample size and n_2 is the second sample size. W is the sum of the ranks for the first sample.

U follows a symmetric distribution centered at $n_1 n_2/2$. For a one-sided test, the null hypothesis of identical distributions is rejected if U is more extreme than a critical value U_a, which is based on sample sizes and the desired significance level (α). If U is small, the rank sum of the first sample is relatively large, and there may be evidence that the distribution of the first population is shifted to the right of that of the second population. If U is large, the opposite holds. For a two-sided test, H_0 is rejected if U is greater than or less than critical values determined by $\alpha/2$. If the null hypothesis is rejected in this case, there is simply evidence of differing locations between the two populations without specifying which has larger magnitude.

As mentioned earlier, the MWU test requires fewer assumptions than the t-test. All that are needed are the independence assumption, which can be satisfied through random selection of samples, and the equal variance assumption. Furthermore, in assigning ranks to observations, it is assumed that observations with the same value are given an average rank from the ranks which are due to be assigned.

A recent paper by Fagerland and Sandvik (2009) demonstrates that the MWU test may be quite nonrobust to departures from equal variances and equal distributions. In fact, based on many Monte Carlo simulations of two-sample MWU tests in different situations, the authors conclude that “the problem is more serious than previously thought.” Feltovich (2003) examines the MWU test’s performance under different combinations of sample sizes, distributions, and variance heterogeneity and finds that the test acts “erratically”. For identical underlying distributions, the MWU test performs reasonably well with regard to Type I error. However, for different distributions and variances, the test is conservative (i.e. overcoverage) when the larger sample has the larger variance and liberal (i.e. undercoverage) when the larger sample has the smaller variance. Another study analyzes comparative power of the t-test and MWU test using real-world data sets (Bridge & Sawilowsky, 1999). It finds that the t-test has slightly greater power under symmetric population distributions while the MWU test has a large power advantage for skewed distributions.

Bootstrap Test

A two-sample bootstrap test is another nonparametric method. While bootstrap procedures may be used in a variety of instances, we are interested in a bootstrap hypothesis test that tests equality of location parameters, in particular the means, for two samples. The premise behind a bootstrap test is that one can determine how extreme a given test statistic is by generating a distribution of the statistic through repeated resampling, or selecting observations from the initial samples at random to form new samples. In each resampling, a new statistic is generated under the assumption that each element of the sample is equally likely and can be drawn multiple times. Together these statistics form a discrete probability distribution, which is meant to approximate the underlying population distribution. This allows for the creation of a rejection region for a bootstrap hypothesis test, and the original test statistic can be analyzed in that context.

The two-sample procedure begins by calculating a parametric statistic from the observed data. The reason for bootstrapping with a parametric statistic, and not simply the difference in means, is an increase in the power of the test (Good, 2004). Here, we use Welch's T to account for possible heterogeneity of variances, as recommended by Efron and Tibshirani (1994). We will denote this as T_0 and it is calculated using the formula for $\mathrm{T_W}$ above. We denote the first sample as X, the second sample as Y, and the pooled sample as Z. Next, both samples are adjusted by subtracting their sample mean and adding the pooled mean to each observation, as shown below:

$$x'_i = x_i - \overline{x} + \overline{z}, \;\; i = 1, 2, \dots n_1$$

$$y'_j = y_j - \overline{y} + \overline{z}, \;\; j = 1, 2, \dots n_2$$

This translation of both samples is necessary so that the populations can be assumed to have a common mean (Efron & Tibshirani, 1994).

Many samples, say B of them, are then drawn with replacement from the adjusted samples, X' and Y'. For example, if the original samples are of size n_1 and n_2 respectively, B samples of size n_1 are taken from X' and B samples of size n_2 are drawn from Y'. This does not simply recreate the same two samples every time due to the fact that each resampling is done with replacement. The new samples can be labeled as X^{*b} and Y^{*b}, where $b = 1, 2, \dots B$. From each new pair of samples, the test statistic of interest, T_b, is calculated, thereby gradually forming a distribution.

$$T_b = \frac{\overline{X}^{*b} - \overline{Y}^{*b}}{\sqrt{\frac{S_X^{2*b}}{n_1} + \frac{S_Y^{2*b}}{n_2}}}, \quad b = 1,2, \dots B$$

Once resampling is completed, an attained significance level for a two-sided test can be calculated based on the frequency with which T_b exceeds T_0 in absolute magnitude. If T_b is rarely more extreme than T_0, there is evidence to reject H_0 that the means of the two samples are the same (Efron & Tibshirani, 1994). The formula for the attained significance level, which is equivalent to a p-value, is as follows:

$$p = \frac{\#(|T_b| \geq \ |T_0|)}{B}$$

The above process is outlined in Efron and Tibshirani (1994).

Permutation Test

The ideas behind the permutation test are similar to those behind the bootstrap. At the outset, two samples are drawn and T_0 is calculated. However, instead of adjusting the samples and then resampling many times, the permutation procedure involves pooling the observations from both samples and permuting. In other words, group labels – indicating membership in the first or second sample – are randomly shuffled so as to create two new samples of the same sizes as before. As in the bootstrap, this is done repeatedly to create a distribution of statistics which follow H_0, which is that both underlying population distributions are the same, and therefore so are their location parameters. If T_0 is extreme in relation to the distribution of t-statistics calculated from permutations, there is evidence that the original samples do not come from the same distributions and their means are different. If the original samples were from equal distributions, we would not expect their statistic to differ much from the permutation statistics.

After computing T_0, the permutation process begins with pooling the two samples, X and Y, so as to obtain $Z = X \cup Y$. Z is then permuted repeatedly, say B times, where in each instance the labels on Z's observations are randomly reassigned to produce Z^*. For each permutation, the first n_1 observations make up the new first sample (X^*) while the remaining n_2 observations make up the new second sample (Y^*). Thus, for a given permutation, X^* and Y^* can be characterized as follows:

$$x_i^{*b} = z_i^{*b}, \quad i = 1, 2, \dots n_1 \quad b = 1,2, \dots B$$

$$y_j^{*b} = z_j^{*b}, \; j = n_1 + 1, n_1 + 2, \dots n_1 + n_2 \quad b = 1,2, \dots B$$

We choose to calculate the test statistic T_b in the same way as in the bootstrap test for each X^* and Y^* pair, although alternative statistics could certainly be used. H_0 implies that all permutations of the pooled observations, and therefore all possible values of T_b, are equally likely with probability $\frac{1}{\binom{N}{n_1}}$, where $N = n_1 + n_2$ (Efron & Tibshirani, 1994). That is, if the two population distributions are equal, permuting should have no effect on the test statistic. But if T_0 is extreme relative to the distribution of T_b, there is evidence that T_0 did not merely happen by chance. Permuting produced different results than the original samples. The p-value calculation is the same as in the bootstrap situation.

The major advantage of the bootstrap and permutation tests is that they require few assumptions. For the bootstrap, it is still critical to have independent samples, but neither the equal variance nor the normality assumptions are necessary. The permutation test, on the other hand, does require the two populations to have equal variances as well as interchangeable observations under H_0 (Good, 2004). The normality assumption, or any assumption of identical distributions, is not needed because the theoretical distribution of the test statistic is not determined a priori, but rather through the bootstrapping or permuting process.

A study by Janssen and Pauls (2005) compares Type I error rates for two-sample bootstrap and permutation tests. In general, the latter outperforms the former. The permutation test is superior in cases of heterogeneity of variance and in heavy-tailed distributions, and performs very well under equal-sized samples. The bootstrap may have an advantage, however, when there are unequal sample sizes, unequal variances, and skewed distributions simultaneously present. Good (2004) finds that in a series of simulations with small, equal sample sizes, the permutation test is "close to exact." In addition, as variances became farther apart, performances of both the permutation and bootstrap tests tends to weaken. Romano (1990) finds the permutation test to be asymptotically exact under equal variances and approximately equal sample sizes.

III. Simulation Setup

Simulations were run from customized code written in R, a statistical software package (www.r-project.org). Our general setup and method of reporting results are based off of those used by Fagerland and Sandvik (2009). In total, three separate sets of code were run, one for each of three distributional scenarios: two normal populations, two gamma populations with equal skewness values, and two gamma populations with unequal skewness values. All three sets of code are displayed in raw form in Appendix 1. Within each set, many combinations of three sample sizes, five theoretical standard deviations, and when applicable, six theoretical skewness values, were specified. In order to calculate Type I error, the null hypothesis of equal population means was always predetermined to be true. For each combination, 10,000 simulations were run in which a sample was drawn from each population. Each of the five tests was run within each simulation and H_0 was either rejected or not rejected based on a 5% significance level, and all rejection regions were two-sided. Thus under every combination of sample sizes, standard deviations, and skewness values, 10,000 simulations of each test were run.

The proportion of the 10,000 simulations in which a given test returned a rejection represents our empirical Type I error rate, or *true significance level*. From here on, we denote the true significance level as α_T, and represent it as a percentage, rather than a proportion so that $0 < \alpha_T < 100$.

For every setup, there were nine sample size combinations: (25, 25), (25, 50), (25, 100), (50, 25), (50, 50), (50, 100), (100, 25), (100, 50), and (100, 100). There were also five standard deviation ratios: 1.00, 1.10, 1.25, 1.50, and 2.00. The larger standard deviation was always associated with the first sample. For the gamma with equal skewness setup, the skewness values for both populations were 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0. For the unequal skewness setup, the pairs of skewness values were (1.0, 0.5), (1.5, 1.0), (2.0, 1.5), (2.5, 2.0), and (3.0, 2.5), with the first population having the larger skewness in every instance. The normal distribution setup did not require the use of skewness values since by definition a normal distribution has skewness of zero.

In a gamma distribution, the mean and variance of the population are not independent. In other words, specifying the population standard deviation influences the population mean. Thus in our simulation setup it was not possible to generate unequal population standard deviations without losing equality of means as a side effect. To counteract this quandary, we made adjustments to each sample by subtracting the population mean from each observation in both the equal skewness and unequal skewness setups. This ensured that the samples behaved as though their population means were the same.

We considered a test to be robust in a given situation if α_T was within a 10% range of the target level, 5. That is, if $4.5 \leq \alpha_T \leq 5.5$. Similarly, we defined somewhat robust as α_T being within a 20% range of the target, or $4.0 \leq \alpha_T \leq 5.0$. If $\alpha_T \leq 4.0$ or $\alpha_T \geq 6.0$, the test was nonrobust. It should be reiterated that not all of our tests have the same assumptions. Therefore some Type I error rates were calculated when no assumptions were being violated. Referring to these results as either “robust” or “nonrobust” would be technically incorrect since robustness by definition requires some breakdown of assumptions. Nonetheless, for simplification and ease of comparison we used the robustness criteria outlined above for all results.

IV. Results

The descriptions and tables below describe only selected results. Full results for each test can be found in Tables A – E in Appendix 2. The color scheme in the tables is as follows: red represents nonrobust undercoverage, yellow represents somewhat robust undercoverage, green represents robustness, blue represents somewhat robust overcoverage, and purple represents nonrobust overcoverage.

Two-Sample Student's t-test

Results for the Student's t-test varied greatly. As might be expected, the test performed best when both the equal variance and normality assumptions were simultaneously satisfied. Under this scenario, α_T ranged between 4.4 and 5.5. Even if only the normality assumption was met, the test was robust if sample sizes were equal, with α_T ranging between 4.5 and 5.5.

When neither the normality nor the equal variance assumption was satisfied, the statistic was fairly robust as long as sample sizes were equal. For sample sizes (25, 25) under both equal and unequal skewness, nonrobustness occurred in only a few instances involving a large skewness value and standard deviation ratio (SDR). Robustness improved marginally as sample sizes increased to (50, 50) and (100, 100).

The t-test performed substantially worse when sample sizes were unequal. When the first sample was smaller than the second, undercoverage was widespread in both the normal and equal skewness cases. α_T inflated as the SDR increased but seemed hardly affected by skewness. This problem worsened as the discrepancy in sample sizes widened, with α_T reaching 18.7 for the (25,100) combination, as seen in Table 1. For this and each table that follows, quantities represent the number of rejections out of 10,000 simulations.

Table 1: Student's t-test, Equal Skewness, (25, 100)

SD Ratio						
2.00	1797	1788	1791	1869	1779	1873
1.50	1173	1199	1248	1226	1207	1193
1.25	863	814	842	835	861	801
1.10	623	591	652	582	576	581
1.00	491	487	459	433	434	444
Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**

Interestingly, for the unequal skewness setup, overcoverage was prevalent for small SDR and skewness values. However, undercoverage was present for large standard deviation or skewness values. The dividing line between the two tended to produce a few robust or somewhat robust results.

The opposite results occurred when the first sample was the larger one. For normal and equal skewness scenarios, α_T deflated with larger SDRs to produce extreme overcoverage. In the worst case – (100, 25) – α_T was as low as 0.4, shown below in Table 2. For the unequal skewness setup, undercoverage was associated with small SDR and skewness values and overcoverage with large SDR and skewness values.

Table 2: Student's t-test, Normal setup, (100, 25)

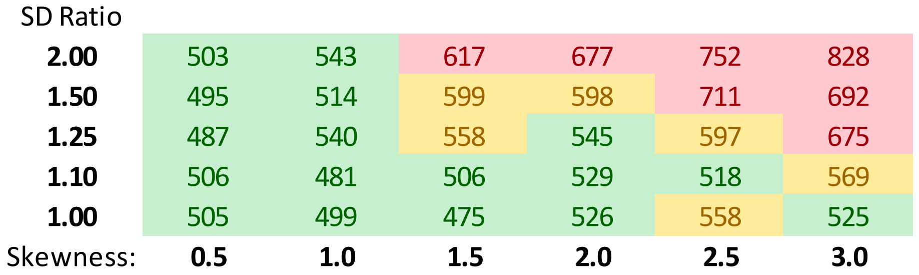
SD Ratio	
2.00	38
1.50	150
1.25	249
1.10	359
1.00	502

Two-Sample Welch's t-test

Without a doubt, Welch's t-test outperformed Student's t-test with respect to Type I error. The test was robust under normal distributions regardless of sample sizes or SDR. Therefore, the test works well when its assumptions are satisfied.

When sample sizes were the same, Welch's t-test behaved similar to Student's t-test, having only a few instances of nonrobustness. This was true in both the equal and unequal skewness scenarios. When the first sample was smallest, α_T inflated into nonrobust values as SDR and skewness increased. However, for skewness values of 1 or below, the test almost always remained robust, as seen in Table 3. As before, the issue of undercoverage intensified as the difference in sample sizes grew. Although nonrobust, α_T was not nearly as large as in the previous case, not exceeding 9.7.

Table 3: Welch's t-test, Equal Skewness, (25, 50)



SD Ratio						
2.00	503	543	617	677	752	828
1.50	495	514	599	598	711	692
1.25	487	540	558	545	597	675
1.10	506	481	506	529	518	569
1.00	505	499	475	526	558	525
Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**

For situations where the first sample was largest, Welch's t-test outperformed Student's t-test significantly. Results tended to be at least somewhat robust, except for a handful of cases, most of which occur under the (100, 25) combination. All nonrobustness was in the direction of undercoverage and occurred for combinations of small SDRs and large skewness values. Even so, the worst case was a α_T of 7.7. For equal skewness, there was no nonrobustness for either

(50, 25) or (100, 50). The latter results are shown in Figure 4. The same could be said for unequal skewness save for a couple cases where α_T barely exceeded 6.0.

Table 4: Welch's t-test, Equal Skewness, (100, 50)

SD Ratio						
2.00	495	547	489	461	510	542
1.50	522	484	475	483	514	448
1.25	482	483	467	514	486	524
1.10	513	437	490	506	489	500
1.00	492	453	531	518	541	526
Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**

Mann-Whitney U test

Results for the MWU test were concerning. Under equal sample sizes, the test was robust for normal distributions except for SDR = 2.0. For equal and unequal skewness, the test was nonrobust in most circumstances. Robustness tended to occur for small SDR and skewness values in the equal skewness case, and was almost always present when SDR = 1.0. For unequal skewness, robustness typically was found when skewness values were small and SDR was 1.25 or 1.50. Outside of these pockets of robustness, undercoverage was extreme for large SDRs or skewness values. Furthermore, the situation worsened as sample sizes increased. For equal skewness, the largest α_T for (25, 25) was 46.9. This increased to 74.5 for (50, 50) and a remarkable 95.5 for (100, 100). Thus the test rejected a true H_0 over 95% of the time. This last result is displayed in Table 5.

Table 5: MWU Test, Equal Skewness, (100, 100)

SD Ratio						
2.00	853	1943	3941	6452	8494	9545
1.50	676	1162	2168	4124	6548	8568
1.25	562	675	1096	2002	3822	6020
1.10	497	536	619	874	1451	2824
1.00	486	516	526	481	487	478
Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**

In cases where the first sample is smaller than the second, α_T inflated more quickly for normal distributions. For equal skewness, results followed the same pattern as for equal sample sizes. Robustness was retained, however, when SDR = 1.00. As shown in Table 6, for unequal skewness, overcoverage appeared when skewness values were small or SDR = 1.25 and was at its worst in the case of (25, 100). Otherwise, undercoverage was the norm.

Table 6: MWU Test, Unequal Skewness, (25, 100)

SD Ratio					
2.00	519	1086	2085	3583	5183
1.50	278	456	601	883	2067
1.25	234	320	341	374	412
1.10	218	316	459	553	847
1.00	193	342	541	858	1427
Skewness:	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**

If the first sample is larger, there was overcoverage under normal distributions for larger SDRs. Under equal skewness, for (50, 25) and (100, 50), approximately half of the results (combinations of lower skewness and SDRs) were robust. The other half involved undercoverage. There was also some overcoverage under equal skewness for (100, 25). When SDR = 1.00, results remained robust. For unequal skewness, undercoverage was present in almost every circumstance. As before, undercoverage was least severe when SDR = 1.25.

Bootstrap Test

The bootstrap test as a whole performed quite well. Under normal distributions it was robust in every situation, except for two where it was somewhat robust. Under equal skewness, as long as skewness was less than or equal to 1.5, the test was never nonrobust and usually was robust. Otherwise, results tended to improve as sample sizes increased. (25, 25), (50, 25), and (50, 50) were all associated with mild overcoverage (α_T between 2.9 and 4.5). On the other hand, (25, 50), (25, 100), and to a lesser extent (100, 25) were all associated with undercoverage. However, α_T did not exceed 7.9. See Table 7. (50, 100), (100, 50), and (100, 100) were robust or somewhat robust in virtually all cases.

Table 7: Bootstrap Test, Equal Skewness, (25, 100)

SD Ratio						
2.00	509	480	518	598	621	790
1.50	495	497	546	642	684	787
1.25	497	488	535	636	717	755
1.10	467	485	539	544	633	709
1.00	510	518	520	579	634	683
Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**

Unequal skewness results were similar and in fact slightly more robust. Robustness was typically maintained for skewness combinations (2, 1.5) and below. The most prevalent overcoverage came with (25, 25), but α_T did not drop below 3.2. Undercoverage was most noticeable in (25, 100), but α_T only reached 7.7. As sample sizes increased, results improved to the point that every SDR – skewness combination was robust for (100, 100), as shown in Table 8.

Table 8: Bootstrap Test, Unequal Skewness, (100, 100)

SD Ratio					
2.00	501	458	500	480	493
1.50	510	502	493	479	489
1.25	486	520	521	470	460
1.10	488	472	460	472	454
1.00	495	486	488	511	483
Skewness:	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**

Permutation Test

The permutation test generally did well. It was virtually always robust and never nonrobust under normal distributions. Under equal skewness, undercoverage occurred in combinations of large SDR and skewness values when the sample sizes were equal or the first sample was smaller. The exception was for (100, 100), where no nonrobustness was present. Undercoverage was worst in (25, 50) and (25, 100), where α_T reached 8.0 and 8.2, respectively. For (50, 25) and (100, 25), overcoverage occurred for large SDR and skewness combinations. However, the problem was considerably more severe for (100, 25), in which α_T dropped to 2.4, as displayed in Table 9. (100, 50) had no nonrobustness.

Table 9: Permutation Test, Equal Skewness, (100, 25)

SD Ratio						
2.00	476	414	397	303	312	239
1.50	437	449	440	417	376	303
1.25	482	457	431	454	414	419
1.10	456	485	456	441	479	473
1.00	503	513	477	538	510	513
Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**

Results under unequal skewness mirrored results under equal skewness and were overall more robust. That is, α_T inflated or deflated at a slower rate and in total there were fewer instances of nonrobustness. See Table 10. The maximum α_T was 7.9 while the minimum was 2.8. As in the bootstrap, (100, 100) produced complete robustness.

Table 10: Permutation Test, Unequal Skewness, (100, 25)

SD Ratio					
2.00	517	439	397	342	282
1.50	486	456	422	390	361
1.25	453	495	468	454	415
1.10	549	500	483	504	457
1.00	520	485	498	533	540
Skewness:	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**

Comparison to the Literature

Student's t-test results were consistent with previous literature. The test was usually robust when sample sizes were equal, even in the presence of some skewness. The equal variance assumption seemed to be much more important for robustness than the normality assumption, as evidenced by the relatively large effect on Type I error of altering the SDR compared to changing skewness. And as some literature describes, robustness decreased as sample sizes became farther apart but at the same time increased as sample sizes grew. Welch's t-test, as advertised, was much more robust to the equal variance assumption than Student's t-test. We did not check for the loss of power mentioned by Cressie and Whitford (1986).

Our results confirmed the troubling findings of others with regard to the MWU test. The test struggled mightily with undercoverage under violations of both normality and homogeneity of variance. Like Feltovich (2003), under normal distributions we found the test to exhibit undercoverage when the larger sample had the smaller variance and overcoverage when the larger sample had the larger variance. However, this pattern was hard to generalize to situations outside of the normal distribution setup, as undercoverage showed up virtually everywhere.

For the bootstrap and permutation tests, we were unable to validate Janssen and Pauls' (2005) claim that the permutation test is generally better. Neither were we able to confirm that the bootstrap is superior in combinations of unequal sample sizes, unequal variances, and skewed distributions. Good's (2004) conclusions about the near-exactness of the permutation test for equal sample sizes appeared correct.

V. Conclusions

This study examines the robustness to Type I error rates of five two-sample hypothesis tests under various conditions. Our results suggest that Welch's t-test, the bootstrap test, and the permutation test perform best. Although each of these tests do have some nonrobustness, it is relatively mild. Type I error never exceeds 10%, nor does it fall below 2.9% in any of the three tests.

Student's t-test performs well under equal sample sizes, but tends to struggle elsewhere. Maximum Type I error is about twice as high in Student's t-test as it is in Welch's t-test and Student's t-test has much more severe overcoverage as well. An unavoidable conclusion of our work is that the MWU test simply should not be utilized in many circumstances. It is easier to talk about the few instances when the test behaves well than the majority of scenarios in which it does not. For the MWU to be valid, we find it essential to have normality or exactly identical distributions. Otherwise, conclusions drawn from conducting a MWU test could be substantially incorrect.

Furthermore, we find no striking difference between the bootstrap and permutation tests in our results. This runs somewhat contrary to our expectations and some opinions expressed in the literature. We initially believed the bootstrap test may outperform the permutation test, but this does not appear to be the case.

There are many possibilities for expanding this study. First, alternate distributions could be examined. Here we have only specified normal and gamma distributions in the underlying populations. However there are many other distributional setups, including some with real-world applications, which could be used in a study like this one. Second, a wider variety of SDRs and skewness values could be included. Our skewness values were not that large – never more than 3 – and only differed by 0.5 in the unequal skewness setup. The gap between standard deviations could also be widened to create larger SDRs. Thus more strain could be placed on the tests' assumptions. Third, other two-sample tests could be analyzed. Our main contribution was adding the bootstrap and permutation tests, which in general are less well-known. But there are other parametric and nonparametric tests which would be interesting to analyze. These include a couple of modifications of the MWU test: the Brunner-Munzel and Fligner-Policello tests. Finally, given our results, it would be crucial for future research to take a close look at the MWU test theoretically. There is a lack of literature on the reasoning behind the test's assumptions, but clearly those assumptions are immensely important.

Acknowledgements

First and foremost, I would like to thank Professor Liam O'Brien for his tremendous support, guidance, and patience over the past year.
I would also like to thank the following for their contributions to this work:
Randy Downer
Jim Scott
Mary Fletcher
The Kleinman family

References

Bridge, P. D., & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: Comparative power of the t-test and wilcoxon rank-sum test in small samples applied research. *Journal of Clinical Epidemiology, 52*(3), 229-235. doi:10.1016/S0895-4356(98)00168-1

Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal, 28*(2), 131-148. doi:10.1002/bimj.4710280202

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap* Chapman & Hall/CRC.

Fagerland, M. W., & Sandvik, L. (2009). The wilcoxon-mann-whitney test under scrutiny. *Statistics in Medicine, 28*(10), 1487-1497. doi:10.1002/sim.3561

Feltovich, N. (2003). Nonparametric tests of differences in medians: Comparison of the Wilcoxon-Mann-Whitney and robust rank-order tests. *Experimental Economics, 6*(3), 273-297. doi:10.1023/A:1026273319211

Good, P. I. (2004). *Permutation, parametric, and bootstrap tests of hypotheses* Springer.

Janssen, A., & Pauls, T. (2005). A monte carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems. *Computational Statistics, 20*(3), 369-383.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*(1), 151.

Munroe, Randall. Null Hypothesis. Retrieved from http://xkcd.com/892/

Posten, H. O., Cheng Yeh, H., & Owen, D. B. (1982). Robustness of the two-sample t-test under violations of the homogeneity of variance assumption. *Communications in Statistics - Theory and Methods, 11*(2), 109-126. doi:10.1080/03610928208828221

Ramsey, P. H. (1980). Exact type 1 error rates for robustness of student's t test with unequal variances. *Journal of Educational Statistics, 5*(4), 337-349.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association, 85*(411), 686-692.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the Mann–Whitney U test. *Behavioral Ecology, 17*(4), 688-690. doi:10.1093/beheco/ark016

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111*(2), 352-360. doi:10.1037/0033-2909.111.2.352

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics with applications* Cengage Learning.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29*(3/4), 350-362.

APPENDIX 1: CODE

Exhibit A: Normal Setup

```
# Simulations of five tests under various sample size, skewness, and SD ratio
combinations
# By Gareth Cleveland, Colby College, 2012-2013
# 3/4/13
# Code for Samples taken from normal distributions

sink("boot.out")

rm(list=ls())

library("parallel")
mc.cores <- detectCores()
cat('cores',mc.cores,'\n')


numrep <- 10000 #repetitions for bootstrap and permutations tests
nrep <- 1000 #number of simulations for each combination

# define vectors to store p-values for each type of test
equalvector <- NULL
unequalvector <- NULL
mannvector <- NULL
tside_bootvector <- NULL
permvector <- NULL

# matrices to store p-values for each type of test
equalmatrix <- matrix(nrow=nrep,ncol=0)
unequalmatrix <- matrix(nrow=nrep,ncol=0)
mannmatrix <- matrix(nrow=nrep,ncol=0)
tside_bootmatrix <- matrix(nrow=nrep,ncol=0)
permmatrix <- matrix(nrow=nrep,ncol=0)

#############################################################################
####
# SIMULATIONS

func <- function(a, b, d)
{

  # simulate data using gamma distribution
  normal.1 <- rnorm(a, mean=4, sd=d) #selects random sample of size a with
mean=4, sd=d
  normal.2 <- rnorm(b, mean=4, sd=4) #selects a second sample of size b with
mean=4, sd=4

  # combine the samples into one vector
  overall <- c(normal.1,normal.2)

  # generate the test statistics by subtracting the group mean and adding
overall mean to each observation
```

```
  stat1 <- normal.1-mean(normal.1)+mean(overall)
  stat2 <- normal.2-mean(normal.2)+mean(overall)

  #### T-Test ####
  equal <- t.test(normal.1, normal.2, var.equal = TRUE) #assuming equal
variances
  t.equal <- equal$p.value #store p-value
  unequal <- t.test(normal.1, normal.2, var.equal = FALSE) #assuming unequal
variances
  t.unequal <- unequal$p.value #store p-value

  #### Mann-Whitney U (Wilcox) Test ####
  mann <- wilcox.test(normal.1, normal.2)
  mwu <- mann$p.value #store p-value

  #### Bootstrap Test ####
  # calculate observed test statistic
  tobs <- (t.test(normal.1, normal.2, var.equal = FALSE))$statistic

  tvec <- rep(0,numrep)

  # resample to get new data
  for (i in 1:numrep)
  {
    newdata1 <- sample(stat1,a,replace=T)
    newdata2 <- sample(stat2,b,replace=T)

    t <- (t.test(newdata1, newdata2, var.equal = FALSE))$statistic #calculate
new test statistic with resampled data
    tvec[i] <- t
  }

  pval <- mean(abs(tvec) >= abs(tobs)) #generate p-value based on difference
between resampled data and original data
  tside_boot <- pval # store p-value

  #### Permutation Test ####
  tvec1 <- rep(0,numrep)

  for (i in 1:numrep)
  {
    permsample <- sample(overall) #rearrange the labels of the pooled
observations
    permsample1 <- permsample[1:a] #take first a observations
    permsample2 <- permsample[a+1:b] #take remaining observations
    tvec1[i] <- (t.test(permsample1, permsample2, var.equal =
FALSE))$statistic #generate t statistic for difference between samples
  }

  pval1 <- mean(abs(tvec1) >= abs(tobs)) #generate p-value based on
difference between permuted data and original data
  perm <- pval1 #store p-value

  return(list(t.equal=t.equal, t.unequal=t.unequal, mwu=mwu,
tside_boot=tside_boot, perm=perm))
```

```
}

for (a in c(25,50,100)) #first group of sample sizes
{
  for (b in c(25,50,100)) #second group of sample sizes
  {
    for (d in c(4,4.4,5,6,8)) #different SDs of first sample
    {

        x <- seq(1,nrep)
        y <- mclapply(x,function(x) func(a,b,d), mc.cores=detectCores())

        # store p-value in appropriate vector
        for (j in 1:nrep)
        {
          equalvector[j] <- y[[j]]$t.equal
          unequalvector[j] <- y[[j]]$t.unequal
          mannvector[j] <- y[[j]]$mwu
          tside_bootvector[j] <- y[[j]]$tside_boot
          permvector[j] <- y[[j]]$perm
        }

        cat('sample size 1 =',a, ' sample size 2 =',b, ' SD ratio =',d/4,
'\n')

        # Append p-value vectors onto appropriate matrix
        equalmatrix <- cbind(equalmatrix,equalvector)
        unequalmatrix <- cbind(unequalmatrix,unequalvector)
        mannmatrix <- cbind(mannmatrix,mannvector)
        tside_bootmatrix <- cbind(tside_bootmatrix,tside_bootvector)
        permmatrix <- cbind(permmatrix,permvector)

    }
  }
}

# column labels indicating which sample size, skewness, SD ratio combination
each set of p-values came from
column.labels  <- c('25.25 sdr1','25.25 sdr1.1','25.25 sdr1.25','25.25
sdr1.5','25.25 sdr2',
                    '25.50 sdr1','25.50 sdr1.1','25.50 sdr1.25','25.50
sdr1.5','25.50 sdr2',
                    '25.100 sdr1','25.100 sdr1.1','25.100 sdr1.25','25.100
sdr1.5','25.100 sdr2',
                    '50.25 sdr1','50.25 sdr1.1','50.25 sdr1.25','50.25
sdr1.5','50.25 sdr2',
                    '50.50 sdr1','50.50 sdr1.1','50.50 sdr1.25','50.50
sdr1.5','50.50 sdr2',
                    '50.100 sdr1','50.100 sdr1.1','50.100 sdr1.25','50.100
sdr1.5','50.100 sdr2',
                    '100.25 sdr1','100.25 sdr1.1','100.25 sdr1.25','100.25
sdr1.5','100.25 sdr2',
                    '100.50 sdr1','100.50 sdr1.1','100.50 sdr1.25','100.50
sdr1.5','100.50 sdr2',
                    '100.100 sdr1','100.100 sdr1.1','100.100
sdr1.25','100.100 sdr1.5','100.100 sdr2')
```

```
# applies the labels to the matrices
colnames(equalmatrix) <- column.labels
colnames(unequalmatrix) <- column.labels
colnames(mannmatrix) <- column.labels
colnames(tside_bootmatrix) <- column.labels
colnames(permmatrix) <- column.labels

# sums the number of significant p-values for each combination
c1 = colSums(equalmatrix <= 0.05)
c2 = colSums(unequalmatrix <= 0.05)
c3 = colSums(mannmatrix <= 0.05)
c4 = colSums(tside_bootmatrix <= 0.05)
c5 = colSums(permmatrix <= 0.05)

save.image(file="runnorm.RData")
```

Exhibit B: Equal Skewness Setup

```
# Simulations of five tests under various sample size, skewness, and SD ratio
combinations
# By Gareth Cleveland, Colby College, 2012-2013
# 3/4/13

sink("boot.out")

rm(list=ls())

library("parallel")
mc.cores <- detectCores()
cat('cores',mc.cores,'\n')


numrep <- 10000 #repetitions for bootstrap and permutations tests
nrep <- 1000 #number of simulations for each combination

# define vectors to store p-values for each type of test
equalvector <- NULL
unequalvector <- NULL
mannvector <- NULL
tside_bootvector <- NULL
permvector <- NULL

# matrices to store p-values for each type of test
equalmatrix <- matrix(nrow=nrep,ncol=0)
unequalmatrix <- matrix(nrow=nrep,ncol=0)
mannmatrix <- matrix(nrow=nrep,ncol=0)
tside_bootmatrix <- matrix(nrow=nrep,ncol=0)
permmatrix <- matrix(nrow=nrep,ncol=0)

############################################################################
####
# SIMULATIONS
```

```
func <- function(a, b, d, e)
{

  # simulate data using gamma distribution
  gamma.1 <- rgamma(a, shape=d, scale=e) #selects random sample of size a
with shape=d, scale=e
  gamma.2 <- rgamma(b, shape=d, scale=1) #selects a second sample of size b
with shape=d, scale=1

  gamma.1 = gamma.1-d*e
  gamma.2 = gamma.2-d

  # combine the samples into one vector
  overall <- c(gamma.1,gamma.2)

  # generate the test statistics by subtracting the group mean and adding
overall mean to each observation
  stat1 <- gamma.1-mean(gamma.1)+mean(overall)
  stat2 <- gamma.2-mean(gamma.2)+mean(overall)

  #### T-Test ####
  equal <- t.test(gamma.1, gamma.2, var.equal = TRUE) #assuming equal
variances
  t.equal <- equal$p.value #store p-value
  unequal <- t.test(gamma.1, gamma.2, var.equal = FALSE) #assuming unequal
variances
  t.unequal <- unequal$p.value #store p-value

  #### Mann-Whitney U (Wilcox) Test ####
  mann <- wilcox.test(gamma.1, gamma.2)
  mwu <- mann$p.value #store p-value

  #### Bootstrap Test ####
  # calculate observed test statistic
  tobs <- (t.test(gamma.1, gamma.2, var.equal = FALSE))$statistic

  tvec <- rep(0,numrep)

  # resample to get new data
  for (i in 1:numrep)
  {
    newdata1 <- sample(stat1,a,replace=T)
    newdata2 <- sample(stat2,b,replace=T)

    t <- (t.test(newdata1, newdata2, var.equal = FALSE))$statistic #calculate
new test statistic with resampled data
    tvec[i] <- t
  }

  pval <- mean(abs(tvec) >= abs(tobs)) #generate p-value based on difference
between resampled data and original data
  tside_boot <- pval # store p-value

  #### Permutation Test ####
  tvec1 <- rep(0,numrep)
  for (i in 1:numrep)
  {
    permsample <- sample(overall) #rearrange the labels of the pooled
observations
    permsample1 <- permsample[1:a] #take first a observations
    permsample2 <- permsample[a+1:b] #take remaining observations
    tvec1[i] <- (t.test(permsample1, permsample2, var.equal =
FALSE))$statistic #generate t statistic for difference between samples
  }

  pval1 <- mean(abs(tvec1) >= abs(tobs)) #generate p-value based on
difference between permuted data and original data
  perm <- pval1 #store p-value

  return(list(t.equal=t.equal, t.unequal=t.unequal, mwu=mwu,
tside_boot=tside_boot, perm=perm))

}

for (a in c(25,50,100)) #first group of sample sizes
{
  for (b in c(25,50,100)) #second group of sample sizes
  {
    for (d in c(16,4,16/9,1,.64,4/9)) #shape parameters (alpha) where
skewness=2/sqrt(alpha)
    {
      for (e in c(1,1.1,1.25,1.5,2)) #scale parameters (beta) where
SD=beta*sqrt(alpha)
      {

          x <- seq(1,nrep)
          y <- mclapply(x,function(x) func(a,b,d,e), mc.cores=detectCores())

          # store p-value in appropriate vector
          for (j in 1:nrep)
          {
          equalvector[j] <- y[[j]]$t.equal
          unequalvector[j] <- y[[j]]$t.unequal
          mannvector[j] <- y[[j]]$mwu
          tside_bootvector[j] <- y[[j]]$tside_boot
          permvector[j] <- y[[j]]$perm
          }

        cat('sample size 1 =',a, ' sample size 2 =',b, ' skewness
=',2/sqrt(d), ' SD ratio =',e, '\n')

        # Append p-value vectors onto appropriate matrix
        equalmatrix <- cbind(equalmatrix,equalvector)
        unequalmatrix <- cbind(unequalmatrix,unequalvector)
        mannmatrix <- cbind(mannmatrix,mannvector)
        tside_bootmatrix <- cbind(tside_bootmatrix,tside_bootvector)
        permmatrix <- cbind(permmatrix,permvector)

      }
    }
  }
```

```
}

# column labels indicating which sample size, skewness, SD ratio combination
each set of p-values came from
column.labels  <- c('25.25 sk.5 sdr1','25.25 sk.5 sdr1.1','25.25 sk.5
sdr1.25','25.25 sk.5 sdr1.5','25.25 sk.5 sdr2',
                    '25.25 sk1 sdr1','25.25 sk1 sdr1.1','25.25 sk1
sdr1.25','25.25 sk1 sdr1.5','25.25 sk1 sdr2',
                    '25.25 sk1.5 sdr1','25.25 sk1.5 sdr1.1','25.25 sk1.5
sdr1.25','25.25 sk1.5 sdr1.5','25.25 sk1.5 sdr2',
                    '25.25 sk2 sdr1','25.25 sk2 sdr1.1','25.25 sk2
sdr1.25','25.25 sk2 sdr1.5','25.25 sk2 sdr2',
                    '25.25 sk2.5 sdr1','25.25 sk2.5 sdr1.1','25.25 sk2.5
sdr1.25','25.25 sk2.5 sdr1.5','25.25 sk2.5 sdr2',
                    '25.25 sk3 sdr1','25.25 sk3 sdr1.1','25.25 sk3
sdr1.25','25.25 sk3 sdr1.5','25.25 sk3 sdr2',

                    '25.50 sk.5 sdr1','25.50 sk.5 sdr1.1','25.50 sk.5
sdr1.25','25.50 sk.5 sdr1.5','25.50 sk.5 sdr2',
                    '25.50 sk1 sdr1','25.50 sk1 sdr1.1','25.50 sk1
sdr1.25','25.50 sk1 sdr1.5','25.50 sk1 sdr2',
                    '25.50 sk1.5 sdr1','25.50 sk1.5 sdr1.1','25.50 sk1.5
sdr1.25','25.50 sk1.5 sdr1.5','25.50 sk1.5 sdr2',
                    '25.50 sk2 sdr1','25.50 sk2 sdr1.1','25.50 sk2
sdr1.25','25.50 sk2 sdr1.5','25.50 sk2 sdr2',
                    '25.50 sk2.5 sdr1','25.50 sk2.5 sdr1.1','25.50 sk2.5
sdr1.25','25.50 sk2.5 sdr1.5','25.50 sk2.5 sdr2',
                    '25.50 sk3 sdr1','25.50 sk3 sdr1.1','25.50 sk3
sdr1.25','25.50 sk3 sdr1.5','25.50 sk3 sdr2',

                    '25.100 sk.5 sdr1','25.100 sk.5 sdr1.1','25.100 sk.5
sdr1.25','25.100 sk.5 sdr1.5','25.100 sk.5 sdr2',
                    '25.100 sk1 sdr1','25.100 sk1 sdr1.1','25.100 sk1
sdr1.25','25.100 sk1 sdr1.5','25.100 sk1 sdr2',
                    '25.100 sk1.5 sdr1','25.100 sk1.5 sdr1.1','25.100 sk1.5
sdr1.25','25.100 sk1.5 sdr1.5','25.100 sk1.5 sdr2',
                    '25.100 sk2 sdr1','25.100 sk2 sdr1.1','25.100 sk2
sdr1.25','25.100 sk2 sdr1.5','25.100 sk2 sdr2',
                    '25.100 sk2.5 sdr1','25.100 sk2.5 sdr1.1','25.100 sk2.5
sdr1.25','25.100 sk2.5 sdr1.5','25.100 sk2.5 sdr2',
                    '25.100 sk3 sdr1','25.100 sk3 sdr1.1','25.100 sk3
sdr1.25','25.100 sk3 sdr1.5','25.100 sk3 sdr2',

                    '50.25 sk.5 sdr1','50.25 sk.5 sdr1.1','50.25 sk.5
sdr1.25','50.25 sk.5 sdr1.5','50.25 sk.5 sdr2',
                    '50.25 sk1 sdr1','50.25 sk1 sdr1.1','50.25 sk1
sdr1.25','50.25 sk1 sdr1.5','50.25 sk1 sdr2',
                    '50.25 sk1.5 sdr1','50.25 sk1.5 sdr1.1','50.25 sk1.5
sdr1.25','50.25 sk1.5 sdr1.5','50.25 sk1.5 sdr2',
                    '50.25 sk2 sdr1','50.25 sk2 sdr1.1','50.25 sk2
sdr1.25','50.25 sk2 sdr1.5','50.25 sk2 sdr2',
                    '50.25 sk2.5 sdr1','50.25 sk2.5 sdr1.1','50.25 sk2.5
sdr1.25','50.25 sk2.5 sdr1.5','50.25 sk2.5 sdr2',
                    '50.25 sk3 sdr1','50.25 sk3 sdr1.1','50.25 sk3
sdr1.25','50.25 sk3 sdr1.5','50.25 sk3 sdr2',
                    '50.50 sk.5 sdr1','50.50 sk.5 sdr1.1','50.50 sk.5
sdr1.25','50.50 sk.5 sdr1.5','50.50 sk.5 sdr2',
                    '50.50 sk1 sdr1','50.50 sk1 sdr1.1','50.50 sk1
sdr1.25','50.50 sk1 sdr1.5','50.50 sk1 sdr2',
                    '50.50 sk1.5 sdr1','50.50 sk1.5 sdr1.1','50.50 sk1.5
sdr1.25','50.50 sk1.5 sdr1.5','50.50 sk1.5 sdr2',
                    '50.50 sk2 sdr1','50.50 sk2 sdr1.1','50.50 sk2
sdr1.25','50.50 sk2 sdr1.5','50.50 sk2 sdr2',
                    '50.50 sk2.5 sdr1','50.50 sk2.5 sdr1.1','50.50 sk2.5
sdr1.25','50.50 sk2.5 sdr1.5','50.50 sk2.5 sdr2',
                    '50.50 sk3 sdr1','50.50 sk3 sdr1.1','50.50 sk3
sdr1.25','50.50 sk3 sdr1.5','50.50 sk3 sdr2',

                    '50.100 sk.5 sdr1','50.100 sk.5 sdr1.1','50.100 sk.5
sdr1.25','50.100 sk.5 sdr1.5','50.100 sk.5 sdr2',
                    '50.100 sk1 sdr1','50.100 sk1 sdr1.1','50.100 sk1
sdr1.25','50.100 sk1 sdr1.5','50.100 sk1 sdr2',
                    '50.100 sk1.5 sdr1','50.100 sk1.5 sdr1.1','50.100 sk1.5
sdr1.25','50.100 sk1.5 sdr1.5','50.100 sk1.5 sdr2',
                    '50.100 sk2 sdr1','50.100 sk2 sdr1.1','50.100 sk2
sdr1.25','50.100 sk2 sdr1.5','50.100 sk2 sdr2',
                    '50.100 sk2.5 sdr1','50.100 sk2.5 sdr1.1','50.100 sk2.5
sdr1.25','50.100 sk2.5 sdr1.5','50.100 sk2.5 sdr2',
                    '50.100 sk3 sdr1','50.100 sk3 sdr1.1','50.100 sk3
sdr1.25','50.100 sk3 sdr1.5','50.100 sk3 sdr2',

                    '100.25 sk.5 sdr1','100.25 sk.5 sdr1.1','100.25 sk.5
sdr1.25','100.25 sk.5 sdr1.5','100.25 sk.5 sdr2',
                    '100.25 sk1 sdr1','100.25 sk1 sdr1.1','100.25 sk1
sdr1.25','100.25 sk1 sdr1.5','100.25 sk1 sdr2',
                    '100.25 sk1.5 sdr1','100.25 sk1.5 sdr1.1','100.25 sk1.5
sdr1.25','100.25 sk1.5 sdr1.5','100.25 sk1.5 sdr2',
                    '100.25 sk2 sdr1','100.25 sk2 sdr1.1','100.25 sk2
sdr1.25','100.25 sk2 sdr1.5','100.25 sk2 sdr2',
                    '100.25 sk2.5 sdr1','100.25 sk2.5 sdr1.1','100.25 sk2.5
sdr1.25','100.25 sk2.5 sdr1.5','100.25 sk2.5 sdr2',
                    '100.25 sk3 sdr1','100.25 sk3 sdr1.1','100.25 sk3
sdr1.25','100.25 sk3 sdr1.5','100.25 sk3 sdr2',

                    '100.50 sk.5 sdr1','100.50 sk.5 sdr1.1','100.50 sk.5
sdr1.25','100.50 sk.5 sdr1.5','100.50 sk.5 sdr2',
                    '100.50 sk1 sdr1','100.50 sk1 sdr1.1','100.50 sk1
sdr1.25','100.50 sk1 sdr1.5','100.50 sk1 sdr2',
                    '100.50 sk1.5 sdr1','100.50 sk1.5 sdr1.1','100.50 sk1.5
sdr1.25','100.50 sk1.5 sdr1.5','100.50 sk1.5 sdr2',
                    '100.50 sk2 sdr1','100.50 sk2 sdr1.1','100.50 sk2
sdr1.25','100.50 sk2 sdr1.5','100.50 sk2 sdr2',
                    '100.50 sk2.5 sdr1','100.50 sk2.5 sdr1.1','100.50 sk2.5
sdr1.25','100.50 sk2.5 sdr1.5','100.50 sk2.5 sdr2',
                    '100.50 sk3 sdr1','100.50 sk3 sdr1.1','100.50 sk3
sdr1.25','100.50 sk3 sdr1.5','100.50 sk3 sdr2',

                    '100.100 sk.5 sdr1','100.100 sk.5 sdr1.1','100.100 sk.5
sdr1.25','100.100 sk.5 sdr1.5','100.100 sk.5 sdr2',
                    '100.100 sk1 sdr1','100.100 sk1 sdr1.1','100.100 sk1
sdr1.25','100.100 sk1 sdr1.5','100.100 sk1 sdr2',
                    '100.100 sk1.5 sdr1','100.100 sk1.5 sdr1.1','100.100
sk1.5 sdr1.25','100.100 sk1.5 sdr1.5','100.100 sk1.5 sdr2',
                    '100.100 sk2 sdr1','100.100 sk2 sdr1.1','100.100 sk2
sdr1.25','100.100 sk2 sdr1.5','100.100 sk2 sdr2',
                    '100.100 sk2.5 sdr1','100.100 sk2.5 sdr1.1','100.100
sk2.5 sdr1.25','100.100 sk2.5 sdr1.5','100.100 sk2.5 sdr2',
                    '100.100 sk3 sdr1','100.100 sk3 sdr1.1','100.100 sk3
sdr1.25','100.100 sk3 sdr1.5','100.100 sk3 sdr2')


# applies the labels to the matrices
colnames(equalmatrix) <- column.labels
colnames(unequalmatrix) <- column.labels
colnames(mannmatrix) <- column.labels
colnames(tside_bootmatrix) <- column.labels
colnames(permmatrix) <- column.labels

# sums the number of significant p-values for each combination
c1 = colSums(equalmatrix <= 0.05)
c2 = colSums(unequalmatrix <= 0.05)
c3 = colSums(mannmatrix <= 0.05)
c4 = colSums(tside_bootmatrix <= 0.05)
c5 = colSums(permmatrix <= 0.05)

save.image(file="run3.RData")
```

Exhibit C: Unequal Skewness Setup

```
# Simulations of five tests under various sample size, skewness, and SD ratio
combinations
# By Gareth Cleveland, Colby College, 2012-2013
# 3/4/13
# Code for samples with unequal skewness

sink("boot.out")

rm(list=ls())

library("parallel")
mc.cores <- detectCores()
cat('cores',mc.cores,'\n')


numrep <- 10 #repetitions for bootstrap and permutations tests
nrep <- 10 #number of simulations for each combination

# define vectors to store p-values for each type of test
equalvector <- NULL
unequalvector <- NULL
mannvector <- NULL
tside_bootvector <- NULL
permvector <- NULL
```

```
# matrices to store p-values for each type of test
equalmatrix <- matrix(nrow=nrep,ncol=0)
unequalmatrix <- matrix(nrow=nrep,ncol=0)
mannmatrix <- matrix(nrow=nrep,ncol=0)
tside_bootmatrix <- matrix(nrow=nrep,ncol=0)
permmatrix <- matrix(nrow=nrep,ncol=0)

#############################################################################
####
# SIMULATIONS

func <- function(a, b, d, e)
{

  # simulate data using gamma distribution
  gamma.1 <- rgamma(a, shape=d[1], scale=e) #selects random sample of size a
with shape=d, scale=e
  gamma.2 <- rgamma(b, shape=d[2], scale=1) #selects a second sample of size
b with shape=d, scale=1

  gamma.1 = gamma.1-d[1]*e
  gamma.2 = gamma.2-d[2]

  # combine the samples into one vector
  overall <- c(gamma.1,gamma.2)

  # generate the test statistics by subtracting the group mean and adding
overall mean to each observation
  stat1 <- gamma.1-mean(gamma.1)+mean(overall)
  stat2 <- gamma.2-mean(gamma.2)+mean(overall)

  #### T-Test ####
  equal <- t.test(gamma.1, gamma.2, var.equal = TRUE) #assuming equal
variances
  t.equal <- equal$p.value #store p-value
  unequal <- t.test(gamma.1, gamma.2, var.equal = FALSE) #assuming unequal
variances
  t.unequal <- unequal$p.value #store p-value

  #### Mann-Whitney U (Wilcox) Test ####
  mann <- wilcox.test(gamma.1, gamma.2)
  mwu <- mann$p.value #store p-value

  #### Bootstrap Test ####
  # calculate observed test statistic
  tobs <- (t.test(gamma.1, gamma.2, var.equal = FALSE))$statistic

  tvec <- rep(0,numrep)

  # resample to get new data
  for (i in 1:numrep)
  {
    newdata1 <- sample(stat1,a,replace=T)
    newdata2 <- sample(stat2,b,replace=T)
    t <- (t.test(newdata1, newdata2, var.equal = FALSE))$statistic #calculate
new test statistic with resampled data
    tvec[i] <- t
  }

  pval <- mean(abs(tvec) >= abs(tobs)) #generate p-value based on difference
between resampled data and original data
  tside_boot <- pval # store p-value

  #### Permutation Test ####
  tvec1 <- rep(0,numrep)

  for (i in 1:numrep)
  {
    permsample <- sample(overall) #rearrange the labels of the pooled
observations
    permsample1 <- permsample[1:a] #take first a observations
    permsample2 <- permsample[a+1:b] #take remaining observations
    tvec1[i] <- (t.test(permsample1, permsample2, var.equal =
FALSE))$statistic #generate t statistic for difference between samples
  }

  pval1 <- mean(abs(tvec1) >= abs(tobs)) #generate p-value based on
difference between permuted data and original data
  perm <- pval1 #store p-value

  return(list(t.equal=t.equal, t.unequal=t.unequal, mwu=mwu,
tside_boot=tside_boot, perm=perm))

}

shape.pairs <- list(c(4,16),c(16/9,4),c(1,16/9),c(.64,1),c(4/9,.64)) #see
loop below

for (a in c(25,50,100)) #first group of sample sizes
{
  for (b in c(25,50,100)) #second group of sample sizes
  {
    for (d in shape.pairs) #pairs of shape parameters (alpha) where
skewness=2/sqrt(alpha); first parameter is for gamma.1, second is for gamma.2
    {
      for (e in c(1,1.1,1.25,1.5,2)) #scale parameters (beta) where
SD=beta*sqrt(alpha)
      {

        x <- seq(1,nrep)
        y <- mclapply(x,function(x) func(a,b,d,e), mc.cores=detectCores())

        # store p-value in appropriate vector
        for (j in 1:nrep)
        {
          equalvector[j] <- y[[j]]$t.equal
          unequalvector[j] <- y[[j]]$t.unequal
          mannvector[j] <- y[[j]]$mwu
          tside_bootvector[j] <- y[[j]]$tside_boot
          permvector[j] <- y[[j]]$perm
        }

        cat('sample size 1 =',a, ' sample size 2 =',b, ' skewness 1
=',2/sqrt(d[1]), ' skewness 2 =',2/sqrt(d[2]), ' SD ratio =',e, '\n')

        # Append p-value vectors onto appropriate matrix
        equalmatrix <- cbind(equalmatrix,equalvector)
        unequalmatrix <- cbind(unequalmatrix,unequalvector)
        mannmatrix <- cbind(mannmatrix,mannvector)
        tside_bootmatrix <- cbind(tside_bootmatrix,tside_bootvector)
        permmatrix <- cbind(permmatrix,permvector)

      }
    }
  }
}

# column labels indicating which sample size, skewness, SD ratio combination
each set of p-values came from
column.labels  <- c('25.25 sk.5,1 sdr1','25.25 sk.5,1 sdr1.1','25.25 sk.5,1
sdr1.25','25.25 sk.5,1 sdr1.5','25.25 sk.5,1 sdr2',
                    '25.25 sk1,1.5 sdr1','25.25 sk1,1.5 sdr1.1','25.25
sk1,1.5 sdr1.25','25.25 sk1,1.5 sdr1.5','25.25 sk1,1.5 sdr2',
                    '25.25 sk1.5,2 sdr1','25.25 sk1.5,2 sdr1.1','25.25
sk1.5,2 sdr1.25','25.25 sk1.5,2 sdr1.5','25.25 sk1.5,2 sdr2',
                    '25.25 sk2,2.5 sdr1','25.25 sk2,2.5 sdr1.1','25.25
sk2,2.5 sdr1.25','25.25 sk2,2.5 sdr1.5','25.25 sk2,2.5 sdr2',
                    '25.25 sk2.5,3 sdr1','25.25 sk2.5,3 sdr1.1','25.25
sk2.5,3 sdr1.25','25.25 sk2.5,3 sdr1.5','25.25 sk2.5,3 sdr2',

                    '25.50 sk.5,1 sdr1','25.50 sk.5,1 sdr1.1','25.50 sk.5,1
sdr1.25','25.50 sk.5,1 sdr1.5','25.50 sk.5,1 sdr2',
                    '25.50 sk1,1.5 sdr1','25.50 sk1,1.5 sdr1.1','25.50
sk1,1.5 sdr1.25','25.50 sk1,1.5 sdr1.5','25.50 sk1,1.5 sdr2',
                    '25.50 sk1.5,2 sdr1','25.50 sk1.5,2 sdr1.1','25.50
sk1.5,2 sdr1.25','25.50 sk1.5,2 sdr1.5','25.50 sk1.5,2 sdr2',
                    '25.50 sk2,2.5 sdr1','25.50 sk2,2.5 sdr1.1','25.50
sk2,2.5 sdr1.25','25.50 sk2,2.5 sdr1.5','25.50 sk2,2.5 sdr2',
                    '25.50 sk2.5,3 sdr1','25.50 sk2.5,3 sdr1.1','25.50
sk2.5,3 sdr1.25','25.50 sk2.5,3 sdr1.5','25.50 sk2.5,3 sdr2',

                    '25.100 sk.5,1 sdr1','25.100 sk.5,1 sdr1.1','25.100
sk.5,1 sdr1.25','25.100 sk.5,1 sdr1.5','25.100 sk.5,1 sdr2',
                    '25.100 sk1,1.5 sdr1','25.100 sk1,1.5 sdr1.1','25.100
sk1,1.5 sdr1.25','25.100 sk1,1.5 sdr1.5','25.100 sk1,1.5 sdr2',
                    '25.100 sk1.5,2 sdr1','25.100 sk1.5,2 sdr1.1','25.100
sk1.5,2 sdr1.25','25.100 sk1.5,2 sdr1.5','25.100 sk1.5,2 sdr2',
                    '25.100 sk2,2.5 sdr1','25.100 sk2,2.5 sdr1.1','25.100
sk2,2.5 sdr1.25','25.100 sk2,2.5 sdr1.5','25.100 sk2,2.5 sdr2',
                    '25.100 sk2.5,3 sdr1','25.100 sk2.5,3 sdr1.1','25.100
sk2.5,3 sdr1.25','25.100 sk2.5,3 sdr1.5','25.100 sk2.5,3 sdr2',

                    '50.25 sk.5,1 sdr1','50.25 sk.5,1 sdr1.1','50.25 sk.5,1
sdr1.25','50.25 sk.5,1 sdr1.5','50.25 sk.5,1 sdr2',
                    '50.25 sk1,1.5 sdr1','50.25 sk1,1.5 sdr1.1','50.25
sk1,1.5 sdr1.25','50.25 sk1,1.5 sdr1.5','50.25 sk1,1.5 sdr2',
                     '50.25 sk1.5,2 sdr1','50.25 sk1.5,2 sdr1.1','50.25
sk1.5,2 sdr1.25','50.25 sk1.5,2 sdr1.5','50.25 sk1.5,2 sdr2',
                     '50.25 sk2,2.5 sdr1','50.25 sk2,2.5 sdr1.1','50.25
sk2,2.5 sdr1.25','50.25 sk2,2.5 sdr1.5','50.25 sk2,2.5 sdr2',
                     '50.25 sk2.5,3 sdr1','50.25 sk2.5,3 sdr1.1','50.25
sk2.5,3 sdr1.25','50.25 sk2.5,3 sdr1.5','50.25 sk2.5,3 sdr2',

                     '50.50 sk.5,1 sdr1','50.50 sk.5,1 sdr1.1','50.50 sk.5,1
sdr1.25','50.50 sk.5,1 sdr1.5','50.50 sk.5,1 sdr2',
                     '50.50 sk1,1.5 sdr1','50.50 sk1,1.5 sdr1.1','50.50
sk1,1.5 sdr1.25','50.50 sk1,1.5 sdr1.5','50.50 sk1,1.5 sdr2',
                     '50.50 sk1.5,2 sdr1','50.50 sk1.5,2 sdr1.1','50.50
sk1.5,2 sdr1.25','50.50 sk1.5,2 sdr1.5','50.50 sk1.5,2 sdr2',
                     '50.50 sk2,2.5 sdr1','50.50 sk2,2.5 sdr1.1','50.50
sk2,2.5 sdr1.25','50.50 sk2,2.5 sdr1.5','50.50 sk2,2.5 sdr2',
                     '50.50 sk2.5,3 sdr1','50.50 sk2.5,3 sdr1.1','50.50
sk2.5,3 sdr1.25','50.50 sk2.5,3 sdr1.5','50.50 sk2.5,3 sdr2',

                     '50.100 sk.5,1 sdr1','50.100 sk.5,1 sdr1.1','50.100
sk.5,1 sdr1.25','50.100 sk.5,1 sdr1.5','50.100 sk.5,1 sdr2',
                     '50.100 sk1,1.5 sdr1','50.100 sk1,1.5 sdr1.1','50.100
sk1,1.5 sdr1.25','50.100 sk1,1.5 sdr1.5','50.100 sk1,1.5 sdr2',
                     '50.100 sk1.5,2 sdr1','50.100 sk1.5,2 sdr1.1','50.100
sk1.5,2 sdr1.25','50.100 sk1.5,2 sdr1.5','50.100 sk1.5,2 sdr2',
                     '50.100 sk2,2.5 sdr1','50.100 sk2,2.5 sdr1.1','50.100
sk2,2.5 sdr1.25','50.100 sk2,2.5 sdr1.5','50.100 sk2,2.5 sdr2',
                     '50.100 sk2.5,3 sdr1','50.100 sk2.5,3 sdr1.1','50.100
sk2.5,3 sdr1.25','50.100 sk2.5,3 sdr1.5','50.100 sk2.5,3 sdr2',

                     '100.25 sk.5,1 sdr1','100.25 sk.5,1 sdr1.1','100.25
sk.5,1 sdr1.25','100.25 sk.5,1 sdr1.5','100.25 sk.5,1 sdr2',
                     '100.25 sk1,1.5 sdr1','100.25 sk1,1.5 sdr1.1','100.25
sk1,1.5 sdr1.25','100.25 sk1,1.5 sdr1.5','100.25 sk1,1.5 sdr2',
                     '100.25 sk1.5,2 sdr1','100.25 sk1.5,2 sdr1.1','100.25
sk1.5,2 sdr1.25','100.25 sk1.5,2 sdr1.5','100.25 sk1.5,2 sdr2',
                     '100.25 sk2,2.5 sdr1','100.25 sk2,2.5 sdr1.1','100.25
sk2,2.5 sdr1.25','100.25 sk2,2.5 sdr1.5','100.25 sk2,2.5 sdr2',
                     '100.25 sk2.5,3 sdr1','100.25 sk2.5,3 sdr1.1','100.25
sk2.5,3 sdr1.25','100.25 sk2.5,3 sdr1.5','100.25 sk2.5,3 sdr2',

                     '100.50 sk.5,1 sdr1','100.50 sk.5,1 sdr1.1','100.50
sk.5,1 sdr1.25','100.50 sk.5,1 sdr1.5','100.50 sk.5,1 sdr2',
                     '100.50 sk1,1.5 sdr1','100.50 sk1,1.5 sdr1.1','100.50
sk1,1.5 sdr1.25','100.50 sk1,1.5 sdr1.5','100.50 sk1,1.5 sdr2',
                     '100.50 sk1.5,2 sdr1','100.50 sk1.5,2 sdr1.1','100.50
sk1.5,2 sdr1.25','100.50 sk1.5,2 sdr1.5','100.50 sk1.5,2 sdr2',
                     '100.50 sk2,2.5 sdr1','100.50 sk2,2.5 sdr1.1','100.50
sk2,2.5 sdr1.25','100.50 sk2,2.5 sdr1.5','100.50 sk2,2.5 sdr2',
                     '100.50 sk2.5,3 sdr1','100.50 sk2.5,3 sdr1.1','100.50
sk2.5,3 sdr1.25','100.50 sk2.5,3 sdr1.5','100.50 sk2.5,3 sdr2',

                     '100.100 sk.5,1 sdr1','100.100 sk.5,1 sdr1.1','100.100
sk.5,1 sdr1.25','100.100 sk.5,1 sdr1.5','100.100 sk.5,1 sdr2',
                     '100.100 sk1,1.5 sdr1','100.100 sk1,1.5 sdr1.1','100.100
sk1,1.5 sdr1.25','100.100 sk1,1.5 sdr1.5','100.100 sk1,1.5 sdr2',
                    '100.100 sk1.5,2 sdr1','100.100 sk1.5,2 sdr1.1','100.100
sk1.5,2 sdr1.25','100.100 sk1.5,2 sdr1.5','100.100 sk1.5,2 sdr2',
                    '100.100 sk2,2.5 sdr1','100.100 sk2,2.5 sdr1.1','100.100
sk2,2.5 sdr1.25','100.100 sk2,2.5 sdr1.5','100.100 sk2,2.5 sdr2',
                    '100.100 sk2.5,3 sdr1','100.100 sk2.5,3 sdr1.1','100.100
sk2.5,3 sdr1.25','100.100 sk2.5,3 sdr1.5','100.100 sk2.5,3 sdr2')


# applies the labels to the matrices
colnames(equalmatrix) <- column.labels
colnames(unequalmatrix) <- column.labels
colnames(mannmatrix) <- column.labels
colnames(tside_bootmatrix) <- column.labels
colnames(permmatrix) <- column.labels

# sums the number of significant p-values for each combination
c1 = colSums(equalmatrix <= 0.05)
c2 = colSums(unequalmatrix <= 0.05)
c3 = colSums(mannmatrix <= 0.05)
c4 = colSums(tside_bootmatrix <= 0.05)
c5 = colSums(permmatrix <= 0.05)

save.image(file="run3.RData")
```

APPENDIX 2: RESULTS

Table A: Student's t-test Results

Sample Sizes	SD Ratio	Rejections out of 10,000												
		Normal		Gamma, Equal Skewness						Gamma, Unequal Skewness				
25, 25	**2.00**	554		510	486	593	589	674	674	490	537	555	553	615
	1.50	504		503	529	521	517	549	578	481	484	487	492	517
	1.25	510		551	519	525	503	485	490	484	488	491	489	459
	1.10	453		529	490	511	476	433	457	549	499	501	445	443
	1.00	539		502	509	445	490	452	427	530	513	475	481	455
			Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
25, 50	**2.00**	1158		1123	1106	1147	1199	1188	1169	483	708	883	968	1021
	1.50	852		851	852	865	852	898	850	324	468	558	617	631
	1.25	698		676	712	684	650	654	669	236	357	443	443	531
	1.10	589		602	562	541	525	501	518	180	299	374	392	404
	1.00	496		482	495	460	465	495	401	163	261	306	370	327
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
25, 100	**2.00**	1773		1797	1788	1791	1869	1779	1873	507	993	1253	1307	1454
	1.50	1183		1173	1199	1248	1226	1207	1193	205	468	625	719	789
	1.25	876		863	814	842	835	861	801	103	271	380	447	515
	1.10	622		623	591	652	582	576	581	68	198	284	311	369
	1.00	498		491	487	459	433	434	444	41	119	190	260	261
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
50, 25	**2.00**	188		179	190	226	243	290	307	555	346	330	312	348
	1.50	240		284	268	272	310	287	340	750	465	413	380	378
	1.25	364		389	359	364	337	340	379	916	680	478	471	473
	1.10	422		438	448	475	442	406	445	1044	793	621	586	587
	1.00	522		515	457	495	502	492	451	1102	886	746	741	618
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
50, 50	**2.00**	480		550	466	558	588	602	623	462	527	503	490	547
	1.50	527		515	503	554	545	533	489	488	493	551	518	513
	1.25	491		494	496	456	478	461	481	526	522	470	512	460
	1.10	525		515	508	511	478	443	502	488	511	523	440	481
	1.00	503		483	484	533	498	454	451	524	509	458	467	495
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
50, 100	**2.00**	1102		1093	1034	1113	1076	1182	1236	483	734	836	924	956
	1.50	888		892	838	872	873	869	862	327	488	555	626	668
	1.25	671		678	730	678	683	706	651	248	375	429	471	529
	1.10	577		591	593	563	557	545	591	192	265	344	387	399
	1.00	493		496	473	475	508	510	501	163	250	309	339	353
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
100, 25	**2.00**	38		59	50	53	69	113	116	555	243	147	149	162
	1.50	150		107	132	149	168	190	179	973	495	315	289	306
	1.25	249		256	279	243	248	270	273	1226	765	575	482	462
	1.10	359		351	368	356	373	366	376	1616	1024	755	748	609
	1.00	502		507	511	479	503	498	473	1834	1202	988	857	742
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
100, 50	**2.00**	174		170	191	187	196	225	276	504	327	297	261	293
	1.50	262		262	261	283	279	296	269	766	518	411	369	388
	1.25	343		333	331	323	380	313	359	885	644	572	499	484
	1.10	406		456	387	408	419	390	419	1005	792	717	598	546
	1.00	437		507	446	521	458	512	476	1126	897	771	705	651
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
100, 100	**2.00**	507		481	492	524	528	574	555	498	472	518	504	517
	1.50	492		510	554	551	522	529	521	510	503	494	488	507
	1.25	514		510	442	513	524	515	524	487	525	521	490	490
	1.10	490		495	499	517	518	496	487	485	477	463	484	484
	1.00	550		488	485	548	504	501	472	505	494	496	531	500
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**

Table B: Welch's t-test Results

Sample Sizes	SD Ratio	Rejections out of 10,000												
		Normal		Gamma, Equal Skewness						Gamma, Unequal Skewness				
	2.00	530		492	470	575	570	655	652	488	529	545	546	598
	1.50	499		498	514	505	502	527	563	470	477	476	477	504
25, 25	**1.25**	505		547	510	511	490	462	468	476	478	477	465	439
	1.10	541		524	488	505	459	413	430	531	494	482	428	416
	1.00	536		498	505	431	476	433	416	513	503	458	468	432
			Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
	2.00	504		503	543	617	677	752	828	513	536	643	727	811
	1.50	523		495	514	599	598	711	692	503	518	545	625	641
25, 50	**1.25**	519		487	540	558	545	597	675	480	486	520	552	635
	1.10	513		506	481	506	529	518	569	474	487	504	516	527
	1.00	507		505	499	475	526	558	525	470	451	489	543	467
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
	2.00	518		526	516	608	706	766	971	556	593	721	786	920
	1.50	496		510	538	604	713	775	881	528	581	639	733	833
25, 100	**1.25**	517		517	518	578	695	754	825	530	560	559	697	771
	1.10	483		483	514	588	585	683	744	528	513	589	637	710
	1.00	498		518	547	550	621	679	723	513	499	526	576	652
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
	2.00	489		505	463	501	458	504	464	558	513	522	489	469
	1.50	485		517	451	473	515	448	428	517	515	475	466	463
50, 25	**1.25**	538		541	510	488	474	464	479	494	531	457	524	528
	1.10	474		504	506	548	505	505	480	526	507	541	554	584
	1.00	512		521	480	507	538	518	551	487	534	557	613	609
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
	2.00	471		544	456	550	578	596	612	461	524	499	485	542
	1.50	521		514	502	546	539	524	483	485	490	546	512	506
50, 50	**1.25**	487		493	494	453	473	457	477	517	520	468	507	455
	1.10	525		511	505	509	475	434	494	478	509	519	434	465
	1.00	502		482	484	531	494	447	439	513	504	457	463	487
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
	2.00	499		492	496	546	577	649	739	492	529	543	616	651
	1.50	494		542	497	543	558	632	622	511	488	501	556	592
50, 100	**1.25**	493		503	576	552	557	558	573	496	487	527	508	588
	1.10	494		511	521	512	541	542	561	489	461	469	501	475
	1.00	492		502	457	475	506	567	549	488	512	496	506	499
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
	2.00	493		520	480	494	439	474	438	543	539	512	516	513
	1.50	505		478	501	543	555	546	512	490	516	546	536	550
100, 25	**1.25**	509		507	510	511	572	589	643	439	547	566	608	636
	1.10	494		477	525	536	558	645	713	540	532	563	631	664
	1.00	492		507	542	543	645	693	773	504	509	570	679	742
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
	2.00	498		495	547	489	461	510	542	481	525	513	460	472
	1.50	493		522	484	475	483	514	448	504	486	454	460	496
100, 50	**1.25**	488		482	483	467	514	486	524	486	473	500	535	541
	1.10	478		513	437	490	506	489	500	473	526	559	546	522
	1.00	446		492	453	531	518	541	526	552	529	562	572	557
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
	2.00	504		477	485	517	527	568	550	498	470	513	504	514
	1.50	490		508	553	547	520	524	521	508	502	493	486	503
100, 100	**1.25**	513		507	442	512	523	513	521	480	524	520	490	483
	1.10	489		494	498	517	515	495	483	483	476	461	480	481
	1.00	549		488	485	548	501	500	471	499	493	493	529	495
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**

Table C: MWU Test Results

Sample Sizes	SD Ratio	Rejections out of 10,000											
		Normal		Gamma, Equal Skewness						Gamma, Unequal Skewness			
25, 25	2.00	630		647	854	1469	2269	3398	4691	511	747	1291	2177	3618
	1.50	546		544	651	906	1425	2244	3264	540	525	555	717	1304
	1.25	488		552	574	672	862	1267	2096	523	564	529	543	547
	1.10	464		525	516	546	550	680	1008	641	583	662	718	1009
	1.00	522		484	493	474	486	460	518	646	694	808	1036	1426
			Skewness:	0.5	1.0	1.5	2.0	2.5	3.0	1.0, 0.5	1.5, 1.0	2.0, 1.5	2.5, 2.0	3.0, 2.5
25, 50	2.00	910		938	1325	2030	3056	4324	5701	498	879	1767	3006	4555
	1.50	699		747	905	1342	1970	3033	4213	384	474	590	817	1685
	1.25	630		631	692	818	1147	1803	2877	363	408	445	413	518
	1.10	566		541	546	565	664	901	1385	363	441	529	657	921
	1.00	503		482	493	467	497	512	463	359	495	649	945	1496
				0.5	1.0	1.5	2.0	2.5	3.0	1.0, 0.5	1.5, 1.0	2.0, 1.5	2.5, 2.0	3.0, 2.5
25, 100	2.00	1090		1203	1613	2377	3566	4848	6285	519	1086	2085	3583	5183
	1.50	793		887	1091	1564	2375	3576	4869	278	456	601	883	2067
	1.25	667		704	775	948	1426	2154	3290	234	320	341	374	412
	1.10	547		570	547	645	722	1086	1653	218	316	459	553	847
	1.00	492		474	475	442	461	488	453	193	342	541	858	1427
				0.5	1.0	1.5	2.0	2.5	3.0	1.0, 0.5	1.5, 1.0	2.0, 1.5	2.5, 2.0	3.0, 2.5
50, 25	2.00	322		412	708	1369	2455	4199	5891	572	687	1290	2598	4420
	1.50	343		432	546	846	1419	2574	4130	679	519	571	707	1391
	1.25	437		457	467	532	790	1336	2297	780	684	588	697	695
	1.10	454		461	475	527	514	709	1059	902	798	828	1019	1375
	1.00	514		505	470	497	530	513	527	953	928	1094	1500	1996
				0.5	1.0	1.5	2.0	2.5	3.0	1.0, 0.5	1.5, 1.0	2.0, 1.5	2.5, 2.0	3.0, 2.5
50, 50	2.00	580		743	1250	2243	3925	5657	7446	503	907	2088	3890	6106
	1.50	533		586	803	1318	2340	3835	5665	523	506	611	926	2149
	1.25	500		533	596	750	1243	2086	3537	588	593	551	564	638
	1.10	512		515	514	586	686	953	1611	644	671	804	1002	1471
	1.00	503		490	468	524	497	491	506	745	793	994	1487	2388
				0.5	1.0	1.5	2.0	2.5	3.0	1.0, 0.5	1.5, 1.0	2.0, 1.5	2.5, 2.0	3.0, 2.5
50, 100	2.00	864		1085	1756	3063	4871	6724	8248	503	1184	2688	4916	7090
	1.50	737		817	1117	1803	3168	4831	6689	409	485	643	1115	2845
	1.25	594		641	780	1057	1696	2872	4635	412	463	443	469	560
	1.10	545		555	578	643	841	1237	2221	437	503	694	1030	1589
	1.00	492		496	459	488	484	535	487	447	657	982	1623	2819
				0.5	1.0	1.5	2.0	2.5	3.0	1.0, 0.5	1.5, 1.0	2.0, 1.5	2.5, 2.0	3.0, 2.5
100, 25	2.00	160		242	524	1170	2680	4698	7097	597	716	1302	2841	5179
	1.50	276		249	406	712	1372	2688	4833	774	570	550	739	1458
	1.25	343		357	390	442	763	1386	2423	916	728	670	695	788
	1.10	399		411	434	440	483	614	1045	1110	973	964	1270	1615
	1.00	472		478	492	480	499	486	502	1253	1154	1291	1742	2377
				0.5	1.0	1.5	2.0	2.5	3.0	1.0, 0.5	1.5, 1.0	2.0, 1.5	2.5, 2.0	3.0, 2.5
100, 50	2.00	337		487	1208	2557	4780	7225	8919	579	994	2356	4988	7514
	1.50	355		456	684	1339	2694	4896	7188	685	590	600	1022	2643
	1.25	395		431	517	696	1308	2522	4489	831	659	677	690	798
	1.10	451		494	466	475	594	905	1823	994	931	1066	1445	2058
	1.00	457		504	442	481	502	526	495	1029	1144	1454	2179	3160
				0.5	1.0	1.5	2.0	2.5	3.0	1.0, 0.5	1.5, 1.0	2.0, 1.5	2.5, 2.0	3.0, 2.5
100, 100	2.00	564		853	1943	3941	6452	8494	9545	566	1415	3590	6610	8815
	1.50	503		676	1162	2168	4124	6548	8568	525	512	764	1322	3846
	1.25	499		562	675	1096	2002	3822	6020	661	567	568	622	779
	1.10	490		497	536	619	874	1451	2824	758	851	1035	1569	2570
	1.00	531		486	516	526	481	487	478	843	1080	1531	2596	4028
				0.5	1.0	1.5	2.0	2.5	3.0	1.0, 0.5	1.5, 1.0	2.0, 1.5	2.5, 2.0	3.0, 2.5

Table D: Bootstrap Test Results

Sample Sizes	SD Ratio	Rejections out of 10,000												
		Normal		Gamma, Equal Skewness						Gamma, Unequal Skewness				
25, 25	**2.00**	531		485	463	527	521	603	590	475	512	500	497	525
	1.50	492		479	500	469	469	463	456	463	452	439	412	395
	1.25	490		541	488	487	446	373	325	470	463	438	393	337
	1.10	446		521	474	472	404	333	327	519	479	451	369	317
	1.00	531		491	490	412	419	353	288	502	476	429	394	331
			Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
25, 50	**2.00**	492		483	514	567	612	670	741	492	499	592	672	749
	1.50	513		481	493	552	544	655	631	489	501	519	580	595
	1.25	507		478	522	530	511	560	629	476	477	486	514	584
	1.10	508		501	471	478	504	479	491	468	473	479	460	456
	1.00	508		499	485	461	484	519	450	474	432	455	484	401
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
25, 100	**2.00**	504		509	480	518	598	621	790	538	520	640	659	774
	1.50	483		495	497	546	642	684	787	508	564	586	658	767
	1.25	507		497	488	535	636	717	755	521	529	526	653	714
	1.10	473		467	485	539	544	633	709	516	504	567	599	674
	1.00	487		510	518	520	579	634	683	517	487	511	539	619
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
50, 25	**2.00**	483		502	470	484	425	442	392	549	499	498	436	387
	1.50	478		515	442	458	467	378	329	502	505	444	411	383
	1.25	530		542	509	462	439	401	391	485	508	439	479	448
	1.10	475		498	489	520	476	435	390	504	494	517	511	521
	1.00	512		518	469	488	511	466	485	478	515	522	585	554
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
50, 50	**2.00**	466		541	448	524	542	553	558	460	512	472	453	492
	1.50	517		510	490	528	530	488	427	485	483	531	482	457
	1.25	486		487	492	451	452	415	395	519	514	451	480	395
	1.10	523		513	506	500	453	398	423	467	506	505	403	407
	1.00	504		480	484	523	474	417	365	504	501	452	438	446
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
50, 100	**2.00**	495		471	477	488	521	558	647	480	503	500	555	593
	1.50	495		532	486	524	514	583	552	515	477	482	515	567
	1.25	484		497	566	531	525	518	536	494	475	501	485	535
	1.10	491		516	514	495	520	518	522	489	455	468	479	448
	1.00	490		494	454	467	490	525	518	483	508	484	474	465
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
100, 25	**2.00**	485		517	476	484	419	437	379	523	526	498	487	458
	1.50	506		466	493	521	530	501	453	476	498	512	515	524
	1.25	511		504	495	500	549	549	601	428	525	527	568	600
	1.10	481		463	500	510	527	607	674	516	496	516	573	615
	1.00	489		494	527	514	606	645	735	481	485	520	621	681
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
100, 50	**2.00**	502		494	543	480	449	484	491	476	523	513	436	441
	1.50	496		512	477	465	467	479	407	498	478	451	448	460
	1.25	489		483	482	459	504	458	466	479	467	481	514	505
	1.10	475		508	434	485	491	466	449	476	521	541	523	485
	1.00	442		491	450	521	500	507	486	550	518	542	548	519
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
100, 100	**2.00**	502		473	481	502	502	533	504	501	458	500	480	493
	1.50	501		507	557	548	507	503	493	510	502	493	479	489
	1.25	518		509	436	508	512	488	495	486	520	521	470	460
	1.10	492		501	502	520	510	484	465	488	472	460	472	454
	1.00	550		492	484	543	489	489	436	495	486	488	511	483
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**

Table E: Permutation Test Results

Sample Sizes	SD Ratio	Rejections out of 10,000												
		Normal		Gamma, Equal Skewness						Gamma, Unequal Skewness				
25, 25	**2.00**	558		514	501	604	607	703	723	497	559	576	569	654
	1.50	506		508	538	526	527	589	626	490	492	507	527	567
	1.25	508		551	522	533	524	520	537	498	505	526	518	511
	1.10	457		517	500	527	497	482	514	554	514	524	484	480
	1.00	535		497	516	456	512	486	490	535	529	492	512	492
			Skewness:	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
25, 50	**2.00**	539		533	567	632	688	746	804	508	544	639	722	789
	1.50	543		508	525	599	588	683	657	489	510	534	588	608
	1.25	527		496	544	551	520	582	640	467	478	498	533	590
	1.10	522		504	484	500	501	491	534	455	476	482	476	498
	1.00	513		505	496	465	500	521	486	463	437	459	518	418
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
25, 100	**2.00**	554		556	525	577	660	660	824	533	566	657	681	772
	1.50	514		519	527	567	640	656	710	492	545	569	619	668
	1.25	529		526	499	520	589	617	623	496	519	484	562	581
	1.10	482		482	478	527	505	520	544	490	481	513	497	524
	1.00	499		509	509	480	518	495	486	473	446	444	439	461
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
50, 25	**2.00**	471		484	445	474	440	464	425	553	489	489	443	431
	1.50	466		497	435	457	471	400	375	531	503	454	433	420
	1.25	532		534	497	464	441	417	429	516	527	443	494	493
	1.10	471		500	489	521	489	462	448	548	515	539	522	548
	1.00	513		519	473	494	516	483	517	517	543	563	595	587
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
50, 50	**2.00**	479		553	462	566	600	624	645	473	534	510	500	573
	1.50	528		518	519	562	566	550	514	488	497	557	535	540
	1.25	496		494	507	465	489	480	514	539	525	486	525	483
	1.10	527		524	511	515	485	462	520	491	519	536	460	503
	1.00	501		483	490	544	512	485	484	531	523	470	484	517
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
50, 100	**2.00**	517		510	511	551	588	638	714	493	531	539	605	636
	1.50	514		548	495	540	557	612	581	507	481	491	536	581
	1.25	501		506	580	541	553	533	546	489	477	512	481	549
	1.10	499		517	525	503	531	525	532	475	448	459	486	452
	1.00	493		499	453	478	492	543	519	478	505	476	478	450
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
100, 25	**2.00**	455		476	414	397	303	312	239	517	439	397	342	282
	1.50	486		437	449	440	417	376	303	486	456	422	390	361
	1.25	488		482	457	431	454	414	419	453	495	468	454	415
	1.10	478		456	485	456	441	479	473	549	500	483	504	457
	1.00	492		503	513	477	538	510	513	520	485	498	533	540
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
100, 50	**2.00**	497		487	538	471	444	480	498	482	507	493	436	440
	1.50	485		505	471	463	468	484	421	508	487	442	440	467
	1.25	482		472	469	452	491	459	480	506	472	492	517	503
	1.10	472		511	428	483	487	471	468	486	527	545	532	492
	1.00	453		490	452	516	503	507	492	558	543	556	553	521
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**
100, 100	**2.00**	509		479	496	531	541	584	568	499	477	523	511	533
	1.50	499		508	561	558	527	538	533	508	503	507	504	523
	1.25	517		515	450	513	529	524	545	495	529	523	492	508
	1.10	490		494	496	521	527	504	502	487	483	467	487	500
	1.00	551		491	492	544	507	511	480	502	500	501	534	518
				0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**1.0, 0.5**	**1.5, 1.0**	**2.0, 1.5**	**2.5, 2.0**	**3.0, 2.5**