A COMPARISON OF THE PERFORMANCE OF SEVERAL SOLUTIONS TO THE BEHRENS-FISHER PROBLEM

by

BARBARA ROSE KUZMAK

B. S., CORNELL UNIVERSITY, 1978
M. S., KANSAS STATE UNIVERSITY, 1982

A MASTERS REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1986

Approved by:

Paul Nelson

Major Professor

ACKNOWLEDGEMENTS

I wish to express my gratitude to Dr. Paul Nelson. His keen guidance, helpful suggestions and hard work made it possible for me to complete my report rapidly.

I am grateful to my committee members, Dr. John Boyer and Dr. Shian Perng for their critical review of this manuscript.

I would like to thank Dr. John Boyer and Dr. George Milliken for finding time to listen. I will always remember their gems of wisdom.

I wish to express my appreciation to the Department of Statistics for an enjoyable two years.

I am indebted to my friends whose support made my stay in Manhattan worth it.

I would like to thank Ms. Faye Roop for her assistance in using SCRIPT.

THIS BOOK CONTAINS NUMEROUS PAGES WITH DIAGRAMS THAT ARE CROOKED COMPARED TO THE REST OF THE INFORMATION ON THE PAGE.

THIS IS AS RECEIVED FROM CUSTOMER.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

INTRODUCTION

Student's pooled t-test is a powerful procedure for testing the equality of two means based on independent random samples $\{X_1\text{ 's}\}$ and $\{X_2\text{'s}\}$ from normal populations with equal variances. If the assumption of normality is violated the test is fairly robust in holding the specified level and maintaining good power. However, the test can perform poorly if $\sigma_1^2 \neq \sigma_2^2$, especially if the sample sizes are very different. As an alternative, consider the statistic:

$$t' = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^{1/2}}$$

In large samples, t' is approximately distributed N(0,1) whether or not the variances are equal. But if n_1 and n_2 are "small," say both less than 30, t' may have a distribution which is quite different from a standard normal.

Testing for differences between two means of normal distributions when the variances are unequal is generally called the Behrens-Fisher problem. Linnik (1966) has showed that there is no reasonable exact solution to this problem. Several approximate solutions exist for this problem. The tests Behrens and Fisher developed about 50 years ago are not frequently used. Behrens' test remains unclear, and Fisher's test lacks uniform size. Approximate tests of Welch and several others are more popular because they are easier to use, and require just some simple calculations and a t-table.

Another procedure which can be used to test for differences of two means is the Wilcoxon rank sum test. This nonparametric test is an exact test of location when the underlying distributions are otherwise identical. Not much is known as to how this test performs as a solution to the Behrens-Fisher problem.

When testing for differences of two means, sometimes the assumptions go unchecked. Consequently, tests are used inappropriately. How well do Student's t-test and Wilcoxon's rank sum test perform if the variances are moderately unequal? extremely unequal? What is the effect of differing sample sizes? How accurate are the results of Welch's approximate t-test when the populations are nonnormal?

I will investigate these questions via a simulation study. I will compare the power of Student's pooled t-test, Welch's approximate t-test and Wilcoxon's rank sum test under a variety of conditions encompassing unequal variances, departures from normality and differing sample sizes.

LITERATURE REVIEW

The problem of testing the equality of two means based on independent random samples from normal distributions with unequal variances has been studied for over 50 years. In this section, I will review some major historical contributions to the Behrens-Fisher problem, especially the works of Fisher and Welch. My discussion includes a brief presentation of several approximate tests based on Fisher's or Welch's solution. Since Student's pooled t-test and Wilcoxon's rank sum test are often used under conditions of unequal variances, I will compare their performance to tests specifically developed for the Behrens-Fisher problem.

Behrens developed the first approximate solution in 1929. Then, in 1935 Fisher used fiducial probability to derive Behrens' solution. Fisher found a distribution based on a weighted difference of two separate t distributions. He showed that the test statistic d:

$$d = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^{1/2}}$$

could be expressed as:

$$d = t_1 \sin R + t_2 \cos R$$

where t_1 and t_2 are independent random variables with t-distributions having n_1 and n_1 degrees of freedom, respectively and $\tan R = (S_1/S_2)(n_2/n_1)^{1/2}$ He found the distribution of d by using his idea of logical inversion. This step made his solution controversial (Robinson 1982). Sukhatme, Fisher and Hearly tabulated significant points to Fisher's solution for $0^\circ \le \theta \le 90^\circ$, where $\theta = arctan(R)$. Fisher and Yates published these tables in 1938. However, these tables did not gain wide acceptance because the test's nominal significance level varied from 3.9% to 5.0% throughout the parameter space (James 1959).

Welch (1947) derived a second order series approximation to the Behrens-Fisher distribution which maintained power under the null hypothesis. Aspin (1948) expanded the series to include fourth order terms. Pearson and Hartley (1954) tabulated critical values to the Welch-Aspin solution for $f_1 > 6$ and $f_2 > 6$ at $\alpha = .05$ and $f_1 > 10$ and $f_2 > 10$ at $\alpha = .05$ where $f_i = n_i - 1$. Some statisticans criticized the Welch-Aspin tables because Welch verified them for only one set of parameters. Wang (1971) and Lee and Gurland (1975) found them to be accurate for several sets of sample sizes. After investigating the nominal significance level of the Welch-Aspin solution, Lee and Gurland (1975) reported the maximum deviation for the examined sample sizes to be .001 from .05.

Welch also produced two approximations to the asymptotic series solution using a Student's t-distribution with degrees of freedom weighted by the sample variances and sizes. In 1947, Welch proposed approximating the distribution of d by a t-distribution whose degrees of freedom are functions of n_1, n_2, S_1^2 and S_2^2. Welch's approximate degrees of freedom are obtained by equating $S_1^2/n_1 + S_2^2/n_2$ with the moments of a scaled chi-square variable (Wang 1971). The estimate of f is:

$$\hat{f} = \frac{(f_1)(f_2)}{f_2\hat{C}^2 + f_1(1 - \hat{C})^2} \qquad (1)$$

where:

$$\hat{C} = \frac{S_1^2/n_1}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

In 1937 Welch had developed another approximate solution which has not received much recognition. Fenstad (1983) compared Welch's two approximate solutions by Monte Carlo simulation. He found that the earlier obscure test maintained its nominal level closer to 5% than the more popular approximation.

Several statisticans compared the power of Fisher's test and Welch's two tests. Scheffé (1970) argued that the Welch-Aspin solution was more powerful than Fisher's test, because for constant conditions the critical values of Fisher's test were larger than Welch-Aspin's test. He also stated that the power of the approximate Welch test was similar to the Welch-Aspin test because the critical points were close in value. Wang (1971) examined the size of these three tests. She found that the Fisher's test deviated the most from a nominal significance level whereas the Welch-Aspin test varied the least. The size of Fisher's test was always below the nominal level; thus, it was more conservative than both of Welch's tests. For $f_1 = 6, f_2 = 12$ and $.031 \leq \sigma_1^2/\sigma_2^2 \leq 32$, Wang (1971) reported the maximum deviation from $\alpha = .05$ for Fisher's test was .0144 below α , whereas it was .0024 for Welch's approximate test. Lee and Gurland (1975) found similar results for $n_1 = 5$ and $n_2 = 9$.

Other statisticans have developed solutions to the Behrens- Fisher problem. Most of these tests are approximations to either Fisher's or Welch's solution. Cochran and Cox (1950), Banerjee (1961) and Patil (1965) devised approximations to Fisher's solution. Banerjee and Cochran and Cox's tests are easy to use. However, the tests' size deviates in a similar magnitude to Fisher's test. Wald (1955), Pagurava (1968) and Lee and Gurland (1975) developed approximations to Welch's solution.

Despite the many tests developed for testing for the equality of means when $\sigma_1^2 \neq \sigma_2^2$, researchers often use Student's pooled t-test or Wilcoxon's rank sum test. Statisticans have investigated the performance of these tests when the variances are unequal. For equal sample sizes, Van der Vaart (1961) reported that Student's pooled t-test was insensitive to small inequalities of variance. He also found that the power under the null hypothesis of the nonparametric test exceeded its nominal value in the range of $-1 \leq q \leq 1$ where $q = (\sigma_1^2 - \sigma_2^2)/(\sigma_1^2 + \sigma_2^2)$, while Student's test did not. In contrast, he noted as opposite trend when $n_1 \neq n_2$. He concluded that Wilcoxon's test performed more poorly than Student's test when $1/2 \leq n_1/n_2 \leq 2$. Wetherill (1960) found that Wilcoxon's test was slightly more robust with regard to unequal variances than Student's test. However, he reported that the size and power of the former test was more affected by variation of the third and fourth moment.

I found only one study which compared Student's pooled t-test or Wilcoxon's rank sum test with a test developed for unequal variances. Murphy (1967) simulated the power of Student's and Welch-Aspin's tests. He reported that Welch-Aspin's test maintained its significance level closer to .05 than Student's test while in the range of $1/16 \leq \sigma_2^2/\sigma_1^2 \leq 16$. When $n_1 = n_2$, he found that Student's test was robust with regard to unequal variances.

There exists a lack of information about the performance of commonly used tests which test for differences of two means versus tests devised for the Behrens-Fisher problem. This motivated my comparative study of Student's pooled t-test, Wilcoxon's rank sum test and Welch's approximate t-test.

METHODS

I simulated the power of Student's pooled t-test, Welch's approximate t-test and Wilcoxon's rank sum test under several combinations of means, variances, sample sizes and population distributions. The mean and variance of population 1 were set equal to 0 and 1, respectively. I set the means of population 2 equal to values which yield noncentrality parameter values δ,

$$\delta = \frac{-\mu_2}{\left(\frac{1}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}}$$

equal to 0, 1, 2, 3, 4 or 6. The variances of population 2 were set equal to 1, 2, 4, 9 or 16. I selected equal and unequal samples of small, medium and large size. I ran the simulation using nine different pairs of n_1 and n_2. I investigated the tests' performance under two distributions, the normal and the double exponential. Overall, I simulated the tests' power for 6•5•9•2 = 540 combinations of μ_2, σ_2^2, n_1, n_2 and distributions.

I wrote a computer program which approximated power based on 1000 Monte Carlo simulations for each combination of parameters. I coded the program in FORTRAN77 and executed it on an IBM 4381-1 computer. The program utilized routines in the International Mathematical and Statistical Libraries (IMSL) to access normal, exponential and binomial random number generators, Student's t probability distribution function and the Wilcoxon rank sum test. I generated double exponential deviates from exponential and binomial deviates using the following equation:

$$D_i = B_i \bullet E_i + (B_i - 1) \bullet E_i + \alpha$$

where:

$D_i \sim$ Double Exp $(\alpha, 2\beta^2)$

$B_i \sim$ Bin $(1,.5)$

$E_i \sim \text{Exp}(\beta)$

$\alpha = 0 \text{ or } \mu_2$

$\beta = (1/2)^{1/2} \text{ or } (\sigma_2^2/2)^{1/2}.$

The program estimated the power by recording the proportion of times the simulated test statistics led to rejection of the null hypothesis in 1000 iterations at $\alpha = .10, .05$ and .01. A two sided alternative was used in all cases.

RESULTS

The reported results are estimated powers based on 1000 iterations. Each test was conducted with nominal (but not always actual) level $\alpha = .05$. I do not include powers calculated with $\alpha = .10$ or $\alpha = .01$ in this discussion. I present the results first under the normal distribution then the double exponential distribution. For each distribution, I report the power of each test under the null hypothesis and under several alternative values of μ_2 (with σ_2^2 fixed).

Table 1 contains 95% confidence intervals for the level of each test at various sample size and variance conditions under the normal distribution. I judge test performance on whether the 95% confidence limit includes $\alpha = .05$. This criteria is not valid for Wilcoxon's rank sum test for small sample sizes. The level of this test is based on a permutation distribution and is not exactly $\alpha = .05$. In the case of $n_1 = n_2 = 4$ and $\sigma_1^2 = \sigma_2^2$, the actual level of the test is .0286. This effect is negligible for all but small sample sizes.

Welch's approximate t-test maintained a level $\alpha = .05$ test more often than Student's pooled t-test or Wilcoxon's rank sum test. Welch's test excluded α from the 95% confidence intervals in five instances. The maximum deviation was .008, whereas the remaining cases barely missed .050.

Student's pooled t-test held its level under the condition of equal sample size for all but two cases. It exceeded the nominal level when the variances were extremely unequal ($\sigma_2^2 = 9$ or $\sigma_2^2 = 16$) and the sample sizes were small ($n_1 = n_2 = 4$). In all but one occasion, it performed very poorly for unequal sample sizes and unequal variances. It erred on the conservative side when I associated the larger sample size with the larger variance. Conversely, the level grossly exceeded $\alpha = .05$ when the smaller sample came from the population with larger variance.

The level of Wilcoxon's rank sum test was close to .050 for most simulations of equal sample size. It departed from α when the sample size was small ($n_1 = n_2 = 4$) and variances were equal or slightly unequal. As previously mentioned, $\alpha = .0286$ is the exact level of Wilcoxon's test under small sample sizes of equal variance. The test also exceeded a level $\alpha = .05$ test for large departures

Table 1. 95% Confidence Interval about the Estimated Level under the Normal Distribution.

Sample Size	Test	σ_2^2/σ_1^2 1	2	4	9	16
$n_1=4$ $n_2=4$	S	(.034,.060)	(.044,.074)	(.049,.079)	(.069,.103)*	(.058,.090)*
$n_1=4$ $n_2=4$	W	(.029,.053)	(.035,.061)	(.038,.066)	(.049,.079)	(.041,.069)
$n_1=4$ $n_2=4$	R	(.023,.045)**	(.024,.046)*	(.027,.051)	(.046,.076)	(.041,.069)
$n_1=12$ $n_2=12$	S	(.031,.057)	(.039,.067)	(.032,.058)	(.043,.072)	(.047,.077)
$n_1=12$ $n_2=12$	W	(.030,.056)	(.037,.065)	(.026,.050)	(.038,.066)	(.038,.066)
$n_1=12$ $n_2=12$	R	(.030,.056)	(.034,.060)	(.031,.057)	(.060,.094)*	(.064,.098)*
$n_1=20$ $n_2=20$	S	(.027,.051)	(.026,.050)	(.045,.074)	(.046,.076)	(.040,.068)
$n_1=20$ $n_2=20$	W	(.027,.051)	(.025,.049)*	(.042,.070)	(.044,.072)	(.037,.064)
$n_1=20$ $n_2=20$	R	(.027,.051)	(.029,.053)	(.048,.078)	(.065,.099)*	(.057,.089)*

NOTE:

* Indicates that the interval does not include $\alpha=.05$.
** The exact level of Wilcoxon's test is .0286.
S is Student's pooled t-test.
W is Welch's approximate t-test.
R is Wilcoxon's rank sum test.

Table 1.cont. 95% Confidence Interval about the Estimated Level under the Normal Distribution.

		σ_2^2/σ_1^2				
Sample Size	Test	1	2	4	9	16
n_1= 4 n_2= 8	S	(.038,.066)	(.016,.036)*	(.005,.017)*	(.014,.033)*	(.003,.015)*
n_1= 4 n_2= 8	W	(.045,.075)	(.030,.054)	(.020,.042)*	(.042,.070)	(.026,.050)
n_1= 4 n_2= 8	R	(.040,.068)	(.019,.039)*	(.011,.027)*	(.023,.045)*	(.015,.033)*
n_1= 8 n_2= 4	S	(.035,.061)	(.063,.097)*	(.105,.147)*	(.132,.176)*	(.169,.219)*
n_1= 8 n_2= 4	W	(.036,.064)	(.042,.070)	(.049,.079)	(.034,.060)	(.052,.084)*
n_1= 8 n_2= 4	R	(.036,.062)	(.051,.081)*	(.077,.113)*	(.088,.126)*	(.104,.144)*
n_1= 5 n_2=15	S	(.036,.062)	(.008,.024)*	(.002,.012)*	(.002,.014)*	(.000,.005)*
n_1= 5 n_2=15	W	(.036,.062)	(.024,.046)*	(.032,.058)	(.026,.050)	(.025,.049)*
n_1= 5 n_2=15	R	(.025,.049)*	(.011,.029)*	(.013,.031)*	(.005,.019)*	(.005,.019)*
n_1=15 n_2= 5	S	(.038,.066)	(.076,.112)*	(.130,.174)*	(.220,.274)*	(.225,.279)*
n_1=15 n_2= 5	W	(.042,.070)	(.037,.065)	(.038,.066)	(.046,.076)	(.038,.066)
n_1=15 n_2= 5	R	(.030,.056)	(.044,.074)	(.064,.098)*	(.099,.139)*	(.084,.122)*
n_1=12 n_2=20	S	(.036,.062)	(.030,.056)	(.018,.038)*	(.007,.023)*	(.005,.019)*
n_1=12 n_2=20	W	(.037,.065)	(.050,.080)	(.040,.068)	(.041,.069)	(.036,.062)
n_1=12 n_2=20	R	(.028,.052)	(.030,.056)	(.027,.051)	(.025,.049)*	(.030,.056)
n_1=20 n_2=12	S	(.030,.054)	(.052,.084)*	(.080,.116)*	(.087,.125)*	(.105,.147)*
n_1=20 n_2=12	W	(.030,.056)	(.030,.056)	(.035,.061)	(.031,.057)	(.029,.053)
n_1=20 n_2=12	R	(.025,.049)*	(.042,.070)	(.064,.098)*	(.064,.098)*	(.076,.112)*

NOTE:

* Indicates that the interval does not include α=.05.

S is Student's pooled t-test W is Welch's approximate t-test R is Wilcoxon's rank sum test.

of variance ($\sigma_2^2 = 9$ or $\sigma_2^2 = 16$) when sample sizes were of medium ($n_1 = n_2 = 12$) or large ($n_1 = n_2 = 20$) size. Like Student's t-test, Wilcoxon's rank sum test performed inadequately for unequal sample sizes. However, its accuracy improved slightly with larger sample sizes. Again, Wilcoxon's test did not maintain a nominal significance level under equal variances in two instances ($n_1 = 5, n_2 = 15$ and $n_1 = 20, n_2 = 12$). I ran seven additional simulations of each violation. The 95% confidence interval included α for all but one simulation, therefore I attributed the initial deviations to random error.

I used two criteria to select tests whose performance I appraised under several alternative hypotheses. First, the test must contain α in the 95% confidence limit about the estimated size. Second, the tests' size must be statistically similar to make comparisons. I evaluated the last condition by testing all pairwise comparisons for equality between two proportions at $\alpha = .05$. Table 2 lists the tests that I compared under various alternative hypotheses.

Appendices 1-9 display the outcome of the tests in Table 2. As expected, Student's pooled t-test performed well when the variances were equal. Welch's approximate t-test and Wilcoxon' rank sum test produced results similar to Student's test but only when sample sizes were equal or unequal sample sizes were large ($n_1 = 12$ and $n_2 = 20$ or vice versa).

In situations of unequal variance, the tests' results depend on equality of sample size. For small samples of equal size ($n_1 = n_2 = 4$), the tests' power were alike except under conditions of large departures of variance ($\sigma_2^2/\sigma_1^2 = 16$) where Wilcoxon's test was more powerful than Welch's test (Appendix 1). However, the tests' results were indistinguishable when sample sizes were equal and of medium ($n_1 = n_2 = 12$) or large ($n_1 = n_2 = 20$) size (Appendices 2 and 3, respectively). For most cases of unequal sample size, the appendices reported power for only Welch's test, because it maintained its level at .05 most frequently (Appendices 4-6). Occasionally Wilcoxon's test held its level and competed against Welch's test. This comparison occurred three times and produced three different results (Appendices 7-9). For $n_1 = 15, n_2 = 5$ and $\sigma_2^2 = 2$, Wilcoxon's test was more powerful (Appendix 7). When $n_1 = 12, n_2 = 20$ and $\sigma_2^2 = 4$ or $\sigma_2^2 = 16$, Welch's test performed better (Appendix 8). In contrast, there was no difference between the two tests for $n_1 = 20, n_2 = 12$ and $\sigma_2^2 = 2$ (Appendix 9).

Table 2. Tests to Evaluate under Alternative Hypotheses for the Normal Distribution.

Sample Size		σ_2^2/σ_1^2				
		1	2	4	9	16
n_1= 4	n_2= 4	S W	S W	S W R[1]	W R	W R
n_1=12	n_2=12	S W R	S W R	S W R	S W	S W
n_1=20	n_2=20	S W R	S R	S W R	S W	S W
n_1= 4	n_2= 8	S W R	W	-	W	W
n_1= 8	n_2= 4	S W R	W	W	W	-
n_1= 5	n_2=15	S W R	-	W	W	-
n_1=15	n_2= 5	S W R	W R	W	W	W
n_1=12	n_2=20	S W R	S W R[2]	W R	W	W R
n_1=20	n_2=12	S W R	W R	W	W	W

NOTE:

[1] S vs. W and W vs. R are the only pairwise comparisons that can be made.

[2] S vs. R is the only pairwise comparison that can be made.

S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxom rank sum test.

Table 3 displays 95% confidence limits about the tests' size under the double exponential distribution for the same variance and sample size combinations. Again, Welch's approximate t-test controlled for level $\alpha = .05$ most frequently. However, it maintained its level more often and deviations were closer to .05 under the normal distribution than the double exponential distribution. Welch's test performed poorly for small samples of equal size ($n_1 = n_2 = 4$). It occasionally erred on the conservative side when the variances were equal regardless of sample size. It did not adequately hold its level at $\alpha = .05$ under large inequalities of variance for several pairs of unequal sample sizes ($n_1 = 8, n_2 = 4; n_1 = 15, n_2 = 5$ and $n_1 = 12, n_2 = 20$).

Student's pooled t-test maintained level $\alpha = .05$ for all but one case when samples were of equal size ($n_1 = n_2 = 20$ and $\sigma_2^2 = 16$). Conversely for unequal sample sizes, it held its level only in situations of equal variances and medium or large sample sizes ($n_1 = 5, n_2 = 15; n_1 = 12, n_2 = 20$ or vice versa for both pairs). These results were similar to the test's performance under the normal distribution.

Wilcoxon's rank sum test controlled its level at .05 for most cases of equal sample size. However under equal variances, the test lacked level $\alpha = .05$ on three occasions. Once when $n_1 = n_2 = 4$; but, as previously noted actually $\alpha = .0286$. For the remaining cases ($n_1 = n_2 = 12$ and $n_1 = 4, n_2 = 8$), I ran additional simulations to investigate these violations. The 95% confidence intervals included $\alpha = .05$ in six of seven simulations. Again, I attributed the original deviations to random error. The test's size exceeded .050 for large inequalities of variance ($\sigma_2^2 = 9$ or $\sigma_2^2 = 16$) when samples were of medium ($n_1 = n_2 = 12$) or large ($n_1 = n_2 = 20$) size. This departure also occurred under the normal distribution; however, these deviations were closer to .050 under the double exponential distribution. The test performed poorly for unequal sample sizes, but improved with increased sample size for $n_1 = 12$ and $n_2 = 20$. The inadequate performance of Wilcoxon's test for unequal sample sizes was similar under both distributions.

Table 4 contains the tests of $\alpha = .05$ and equality of levels that I evaluated for the double exponential distribution and Appendices 10-18 display their performance under several alternative hypotheses. Unlike the results obtained when sampling from the normal distribution, there did not exist a test that was most powerful under conditions of equal variance. The tests' performance was

Table 3. 95% Confidence Interval about the Estimated Level under the Double Exponential Distribution.

Sample Size	Test	σ_2^2/σ_1^2 1	2	4	9	16
n_1= 4 n_2= 4	S	(.036,.062)	(.049,.079)	(.029,.053)	(.041,.069)	(.031,.057)
n_1= 4 n_2= 4	W	(.021,.043)*	(.037,.064)	(.014,.032)*	(.024,.048)*	(.015,.035)*
n_1= 4 n_2= 4	R	(.019,.041)**	(.027,.051)	(.019,.041)*	(.037,.065)	(.030,.056)
n_1=12 n_2=12	S	(.027,.051)	(.032,.058)	(.037,.065)	(.042,.070)	(.036,.062)
n_1=12 n_2=12	W	(.026,.050)	(.030,.056)	(.033,.059)	(.030,.056)	(.028,.052)
n_1=12 n_2=12	R	(.024,.046)*	(.032,.058)	(.038,.066)	(.053,.085)*	(.050,.080)*
n_1=20 n_2=20	S	(.034,.060)	(.041,.069)	(.034,.060)	(.048,.078)	(.022,.044)*
n_1=20 n_2=20	W	(.033,.059)	(.039,.067)	(.032,.058)	(.044,.074)	(.019,.041)*
n_1=20 n_2=20	R	(.033,.059)	(.042,.070)	(.042,.070)	(.053,.085)*	(.052,.083)*

NOTE:

* Indicates that the interval does not include α=.05.
** The exact level of Wilcoxon's test is .0286.
S is Student's pooled t-test.
W is Welch's approximate t-test.
R is Wilcoxon's rank sum test. .

Table 3.cont. 95% Confidence Interval about the Estimated Level under the Double Exponential Distribution.

		σ_2^2/σ_1^2				
Sample Size	Test	1	2	4	9	16
$n_1= 4$ $n_2= 8$	S	(.017,.037)*	(.020,.042)*	(.014,.032)*	(.007,.021)*	(.005,.019)*
$n_1= 4$ $n_2= 8$	W	(.015,.033)*	(.030,.054)	(.036,.062)	(.030,.056)	(.030,.056)
$n_1= 4$ $n_2= 8$	R	(.019,.041)*	(.027,.051)	(.031,.057)	(.019,.040)*	(.021,.043)*
$n_1= 8$ $n_2= 4$	S	(.021,.043)*	(.052,.084)*	(.086,.124)*	(.099,.139)*	(.122,.166)*
$n_1= 8$ $n_2= 4$	W	(.024,.046)*	(.031,.057)	(.033,.059)	(.018,.038)*	(.024,.046)*
$n_1= 8$ $n_2= 4$	R	(.028,.052)	(.051,.081)*	(.067,.101)*	(.070,.106)*	(.077,.113)*
$n_1= 5$ $n_2=15$	S	(.035,.061)	(.011,.029)*	(.000,.008)*	(.001,.009)*	(.000,.003)*
$n_1= 5$ $n_2=15$	W	(.031,.057)	(.032,.058)	(.024,.048)*	(.029,.053)	(.026,.050)
$n_1= 5$ $n_2=15$	R	(.030,.054)	(.018,.038)*	(.001,.026)*	(.004,.016)*	(.002,.012)*
$n_1=15$ $n_2= 5$	S	(.041,.069)	(.074,.110)*	(.118,.160)*	(.178,.228)*	(.202,.254)*
$n_1=15$ $n_2= 5$	W	(.037,.064)	(.027,.051)	(.025,.049)*	(.020,.042)*	(.020,.042)*
$n_1=15$ $n_2= 5$	R	(.033,.059)	(.044,.074)	(.057,.089)*	(.070,.104)*	(.068,.102)*
$n_1=12$ $n_2=20$	S	(.037,.065)	(.022,.044)*	(.019,.039)*	(.009,.025)*	(.007,.021)*
$n_1=12$ $n_2=20$	W	(.037,.065)	(.038,.066)	(.035,.061)	(.037,.065)	(.021,.043)*
$n_1=12$ $n_2=20$	R	(.034,.060)	(.029,.053)	(.031,.057)	(.027,.051)	(.024,.048)*
$n_1=20$ $n_2=12$	S	(.028,.052)	(.052,.083)*	(.074,.110)*	(.109,.151)*	(.100,.140)*
$n_1=20$ $n_2=12$	W	(.025,.049)*	(.033,.059)	(.038,.066)	(.044,.072)	(.032,.058)
$n_1=20$ $n_2=12$	R	(.029,.053)	(.044,.074)	(.059,.091)*	(.090,.128)*	(.076,.112)*

NOTE:

* Indicates that the interval does not include α=.05.

S is Student's pooled t-test W is Welch's approximate t-test R is Wilcoxon's rank sum test.

Table 4. Tests to Evaluate under Alternative Hypotheses for the Double Exponential Distribution.

Sample Size	σ_2^2/σ_1^2				
	1	2	4	9	16
n_1= 4 n_2= 4	S	S W	S	S R	S R
n_1=12 n_2=12	S W R	S W R	S W R	S W	S W R[1]
n_1=20 n_2=20	S W R	S W R	S W R	S W	-
n_1= 4 n_2= 8	R	W R	W R	W	W
n_1= 8 n_2= 4	R	W	W	-	-
n_1= 5 n_2=15	S W R	W	-	W	W
n_1=15 n_2= 5	S W R	W R[2]	-	-	-
n_1=12 n_2=20	S W R	W R	W R	W R	-
n_1=20 n_2=12	S R	W R	W	W	W

NOTE:

[1] S vs. R and S vs. W are the only pairwise comparisons that can be made.

[2] no comparison between W and R can be made.

S is the Student pooled t-test.
W is the Welch approximate t-test.
R is Wilcoxon rank sum test.

strongly related to sample size. For small samples of equal size ($n_1 = n_2 = 4$) only Student's pooled t-test held its level at $\alpha = .05$ (Appendix 10). There was no difference among the tests for medium sample size ($n_1 = n_2 = 12$), except for one instance when Welch's test performed more poorly (Appendix 11). Wilcoxon's test performed better for large samples of equal size ($n_1 = n_2 = 20$) (Appendix 12). The results for unequal sample size were similar to the tests' performance under unequal sample sizes, except for small samples ($n_1 = 4, n_2 = 8$ and vice versa) (Appendices 13-18). In this case, only Wilcoxon's test maintained its level at the value $\alpha = .05$ (Appendices 13-14). Welch's test did not perform well under equal variances regardless of sample size.

The results varied under the condition of unequal variances. Only Student's test controlled its level at $\alpha = .05$ under all departures of variance for small samples of equal size. Occasionally Welch's test and Wilcoxon's test competed against Student's test. However, in this case Wilcoxon's test sometimes performed worse than Student's test (Appendix 10). For equal samples of medium ($n_1 = n_2 = 12$) or large ($n_1 = n_2 = 20$) size, the power of Wilcoxon's test frequently surpassed that of Student's and Welch's test (Appendices 11 and 12). Once again, the appendices listed Welch's test most often under conditions of unequal variances and sample sizes. For about half of these entries, Wilcoxon's test competed against Welch's test. When sample sizes were small ($n_1 = 4, n_2 = 8$ or vice versa) or medium ($n_1 = 5, n_2 = 15$ or vice versa), there was no difference between the tests' performance (Appendices 13-16). However, Wilcoxon's test was more powerful than Welch's test for large sample sizes ($n_1 = 12, n_2 = 20$ or vice versa) (Appendices 17 and 18). I did not observe this trend under the normal distribution.

DISCUSSION

I will first discuss the tests' performance under the normal and then under the double exponential distribution. Under each distribution, I first consider equal sample sizes, then unequal sample sizes. I will make test recommendations throughout this section.

With equal sample size from two independent normal distributions, Student's pooled t-test and Welch's approximate t-test closely maintain a level of $\alpha = .05$ and have similar power. It appears from my simulations that they are both insensitive to inequalities of variance. However, Wilcoxon's test is not (Table 5). Van der Vaart (1961) also found that Wilcoxon's test exceeded level $\alpha = .05$ for equal sample sizes and large departures of variance. Therefore, in this case I recommend either Welch's test or the pooled t-test.

For equal sample sizes, Welch's test statistic and the pooled t-test statistic are equal. The difference between the tests in this case is the loss in degrees of freedom for Welch's test when the sample variances are unequal. Table 6 presents the $\alpha = .05$ critical points for both tests as a function of sample size and ratio of variances. Note that the critical values are quite close in most cases. One exception occurs when $n_1 = n_2 = 4$ and $S_2^2/S_1^2 = 9$ or $S_2^2/S_1^2 = 16$. For these two cases, the confidence interval around the nominal significance level of Student's test does not include $\alpha = .05$. Appendices 19-23 contain figures which show the connection between the ratio of variances and the approximate critical points of Welch's test for samples of size 5, 10, 15, 20 or 25.

My results show that Welch's test performs well under conditions of unequal variance for small sample sizes ($n_1 = n_2 = 4$). Murphy (1967) simulated this sample size under the same inequalities of variance and produced results similar to my findings. However, he advocated using caution when sample sizes were less than six, because Welch's t-test could not be verified against the Welch-Aspin series approximation.

When sample sizes are unequal, Student's test is a powerful test to use under conditions of equal variances. However, Student's test and Wilcoxon's test are sensitive to unequal variances and are

Table 5. Recommendations for Testing Equality of Two Population Means Under Various Combinations of Sample Size, Variance and Population Distribution.

	Normal					Double Exponential				
	σ_2^2/σ_1^2					σ_2^2/σ_1^2				
Sample Size	1	2	4	9	16	1	2	4	9	16
$n_1= 4$ $n_2= 4$	SW	SW	SW	SW	SW	S	S	S	S	S
$n_1=12$ $n_2=12$	SW	SW	SW	SW	SW	SR	SR	SR	SR	SR
$n_1=20$ $n_2=12$	SW	SW	SW	SW	SW	R	R	R	R	R
$n_1= 4$ $n_2= 8$	S	W	W	W	W	R	W	W	W	W
$n_1= 8$ $n_2= 4$	S	W	W	W	W	R	W	W	W	W
$n_1= 5$ $n_2=15$	S	W	W	W	W	R	W	W	W	W
$n_1=15$ $n_2= 5$	S	W	W	W	W	R	W	W	W	W
$n_1=12$ $n_2=20$	S	W	W	W	W	R	WR	WR	WR	WR
$n_1=20$ $n_2=12$	S	W	W	W	W	R	WR	WR	WR	WR

NOTE:

S is the Student pooled t-test.
W is the Welch approximated t-test.
R is the Wilcoxon rank sum test.

Table 6. Comparison of the Critical α=.05 t Values Between Student's Pooled t-Test and Welch's Approximate t-Test.

	Student	Welch				
		S_2^2/S_1^2				
Sample Size	Table Value	1	2	4	9	16
n_1= 4 n_2= 4	2.45*	2.45	2.57	2.78	3.18	3.18
n_1=12 n_2=12	2.07	2.07	2.09	2.12	2.16	2.18
n_1=20 n_2=20	1.96	1.96	1.96	2.05	2.07	2.08
n_1= 4 n_2= 8	2.23	2.45	2.31	2.26	2.26	2.31
n_1= 8 n_2= 4	2.23	2.45	2.78	3.18	3.18	3.18
n_1= 5 n_2=15	2.10	2.45	2.26	2.14	2.11	2.11
n_1=15 n_2= 5	2.10	2.45	2.57	2.78	2.78	2.78
n_1=12 n_2=20	1.96	2.07	2.05	2.05	2.06	2.07
n_1=20 n_2=12	1.96	2.07	2.11	2.15	2.18	2.20

NOTE:

*The 95% confidence interval about the approximated power of S under the null hypothesis did not include .05 when σ_2^2/σ_1^2=9 or σ_2^2/σ_1^2=16.

not recommended when $\sigma_1^2 \neq \sigma_2^2$. I suggest the use of Welch's test. Chand (1950) reported this same trend for Student's test. Table 6 shows strong differences between the critical values of Welch's test and Student's t-test when the smaller sample is associated with the larger variance. For this arrangement of sample size and variance, the critical points of Welch's test are highly related to inequalities of variance. For given sample means, sample variances and sample sizes, the test statistic for Student's test is larger than Welch's test. This result coupled with a smaller critical value causes Student's test to frequently exceed .05. When I reverse the order between sample size and variance, the critical points of Welch's test are not strongly dependent on the variances. In this situation, the degrees of freedom are not monotonic decreasing for all departures from equal variances. However, for given sample conditions Student's test statistic is much smaller than Welch's test statistic which results in the frequent exclusion of $\alpha = .05$ from the 95% confidence intervals about the estimated level of the test. I do not have an explanation for the similar performance of Wilcoxon's test under equal versus unequal sample sizes.

The tests' performance under the double exponential distribution is more dependent on sample size than it is under the normal distribution. However, equality of sample size remains important. For small samples of equal size ($n_1 = n_2 = 4$), I recommend using Student's pooled t-test. The test seems to maintain $\alpha = .05$ without regard to variance. I do not suggest using Welch's approximate t-test for small samples because it does not maintain its power under the null hypothesis. As previously noted, Murphy (1967) cautioned its use when sample sizes were less than six. It is difficult to make a recommendation for medium sample sizes ($n_1 = n_2 = 12$). Student's test controls its level at .05 for all variances, whereas the nominal significance level of Wilcoxon's test hardly exceeds .05 in one case. However, Wilcoxon's test performs better than Student's test in 6 out of 20 simulations. There is no difference between these tests for the remaining 12 comparisons. The robustness of Student's test under conditions of nonnormality allows it to rival Wilcoxon's test when samples are of medium size. However for larger sample sizes ($n_1 = n_2 = 20$), I suggest using Wilcoxon's test. Although its nominal significance level barely exceeds .05 on two counts, its power exceeds Welch's and Student's tests for 11 out of 15 simulations. None of the above tests display sensitivity to unequal variances when sample sizes are equal.

For unequal samples of small (n_1 = 4, n_2 = 8 or vice versa) or medium (n_1 = 5, n_2 = 15 or vice versa) size, I recommend Wilcoxon's test when the variances are equal. For unequal variances, I suggest Welch's test for the sole reason that it controls its significance level at .05 most often. In comparisons among the tests at these sample sizes, Welch's test holds α at .05 in 12 out of 20 runs. In contrast, Wilcoxon's test and Student's test maintain a nominal significance level in 6 out of 20 simulations and 2 out of 20 comparisons, respectively. For large sample sizes (n_1 = 12, n_2 = 20 or vice versa), I suggest Wilcoxon's test when variances are equal. For unequal variances, trade-offs exist between Wilcoxon's test and Welch's test. Welch's test controls the nominal significance level more often than Wilcoxon's test. The former test maintains level α = .05 for 8 out of 10 runs, whereas the latter test holds this level in 6 out of 10 simulations. This may not seem different, but in cases of large variances, the level of Wilcoxon's test deviates from α = .05 quite drastically. In regard to power under alternative hypotheses, Wilcoxon's test performs better than Welch's test in 7 out of 20 comparisons. There is no difference between the remaining 13 simulations. Overall, Welch's test is a more conservative recommendation for these sample conditions. Although I advocate using Welch's test in situations of unequal sample sizes and variances, it does not perform as well under the double exponential distribution as under the normal distribution.

CONCLUSIONS

The test to recommend for testing the equality of two population means is highly dependent on what distribution the samples come from and the sample sizes. Equality or inequality of sample size governs whether or not equality of variances is important. If the sample sizes are equal, then equality of variances is not crucial. Under the normal distribution, Student's pooled t-test or Welch's approximate t-test perform well in terms of the nominal significance level and power. In contrast, for unequal sample sizes inequalities of variance are critical. For equal variances, I suggest the use of Student's pooled t-test. However for unequal variances, Welch's approximate t-test is the only test to suggest regardless of the magnitude of the inequality.

Test selection is more complicated when samples come from a nonnormal distribution such as the double exponential distribution. The size of the sample along with its equality or inequality strongly influences test recommendation. Once again, if samples sizes are equal, then departures from equal variance are not important. I recommend Student's test for samples of small size, either Student's test or Wilcoxon's test for medium size and Wilcoxon's test for large sample sizes. For unequal sample sizes, I advocate the use of Wilcoxon's rank sum test when the variances are equal. If the variances are unequal, then I suggest Welch's test for sample sizes of small and medium size and either Welch's or Wilcoxon's tests for large sample sizes. Welch's test does not maintain a nominal significance level as well under a nonnormal distribution as it does under a normal distribution. In general, if either the ratio of σ_2^2/σ_1^2 or distribution is unknown, then I recommend Student's test when the sample sizes are equal and Welch's test when the sample sizes are unequal.

LITERATURE CITED

Aspin, A. A. (1948), "An Examination and Further Developemnt of a Formula Arising in the Problem of Comparing Two Mean Values," *Biometrika*, 35, 88-96.

Aucamp, D. C. (1986), "A Test for the Difference of Means," *Journal of Statistical Computer Simulation*, 24, 33-46.

Banerjee, S. K. (1961), "On Confidence Interval for Two-Means Problem Based on Separate Estimates of Variances and Tabulated Values of t- Table," *Sankhya*, Ser. A, 23, 359-378.

Chand, U. (1950), "Distributions Related to Comparison of Two Means and Two Regression Coefficients," *Annals of Mathematical Statistics* , 21, 507-522.

Cochran, W. G. and Cox, G. M. (1950), *Experimental Designs* , New York: John Wiley.

Fenstand, G. U. (1983), "A Comparison between the U and V Tests in the Behrens-Fisher Problem," *Biometrika*, 70, 300-302.

Fisher, R. A. and Yates, F. (1957), *Statistical Tables for Biological, Agricultural and Medical Research*, 5 th ed. Edinburgh: Oliver and Boyd, Ltd.

James, G. S. (1959), "The Behrens-Fisher Distribution and Weighted Means," *Journal of the Royal Statistical Society*, Ser. B, 21, 73-90.

Lee, A. F. S. and Gurland, J. (1975), "Size and Power of Tests for Equality of Means of Two Normal Populations with Unequal Variances," *Journal of the American Statistical Association*, 66, 605-608.

Linnik, Y. V. (1966), "Latest Investigation on Behrens-Fisher Problem," *Sankhya*, Ser. A, 28, 15-24.

Pagurava, V. I. (1968), "On a Comparison of Means of Two Normal Samples," *Theory of Probability and Its Applications*, 13, no. 3, 527-534.

Patil, V. H. (1955), "Approximation to the Behrens-Fisher Distributions," *Biometrika*, 52, 267-271.

Pearson, E. S. and Hartley, H. O., (eds.), (1954), *Biometrika Tables for Statisticans*, vol. 1, Cambridge: Cambridge University Press.

Robinson, G. K. (1976), "Properties of Student's t and of the Behrens-Fisher Solution to the Two Means Problem," *Annals of Statistics*, 4, 963-971.

________ (1982), "Behrens-Fisher Problem," in *Encyclopedia of Statistical Sciences*, vol. 1, eds. S. Kotz and N. I. Johnson, New York: John Wiley.

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

Van der Vaart, H. R. (1969), "On Robustness of Wilcoxon's Two Sample Test," in *Quantative Methods in Pharmacology*, ed. H. de Jonge New York: Interscience, pp 140-158.

Wald, A. (1955), "Testing the Difference Between the Means of Two Normal Populations with Unknown Standard Deviations," in *Selected Papers in Statistics and Probability by Abraham Wald*, ed. T. W. Anderson, New York: McGraw-Hill, pp 527-534.

Wang, Y. Y. (1971), "Probabilities of the Type I Errors of the Welch Tests for the Behrens-Fisher Problem," *Journal of the American Statistical Association*, 66, 605-608.

Welch, B. L. (1947), "The Generalization of Student's Problem when Several Different Population Variances are Involved," *Biometrika*, 34, 28-35.

APPENDIX

Appendix 1. Comparison of Power for $n_1 = n_2 = 4$ under the Normal Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.125,.169)A	(.375,.435)A	(.662,.720)A	(.890,.926)A	(1.00,1.00)A
1	W	(.107,.149)A	(.329,.389)A	(.626,.684)A	(.861,.901)A	(1.00,1.00)A
1	R	-	-	-	-	-
2	S	(.123,.167)A	(.379,.439)A	(.658,.716)A	(.887,.923)A	(.997,1.00)A
2	W	(.108,.142)A	(.331,.391)A	(.590,.650)B	(.859,.8-9)A	(.987,.998)A
2	R	-	-	-	-	-
4	S	(.131,.175)A[1]	(.360,.420)A[1]	(.657,.715)A[1]	(.882,.920)A[1]	(.994,1.00)A[1]
4	W	(.105,.147)AA	(.307,.365)AA	(.574,.634)BA	(.799,.847)BA	(.981,.995)AA
4	R	(.089,.127) A	(.260,.316) A	(.534,.596) A	(.771,.821) A	(.979,.993) A
9	S	-	-	-	-	-
9	W	(.107,.149)A	(.283,.341)A	(.582,.642)A	(.785,.833)A	(.971,.989)A
9	R	(.107,.149)A	(.307,.365)A	(.618,.678)A	(.820,.866)A	(.980,.994)A
16	S	-	-	-	-	-
16	W	(.103,.143)A	(.290,.348)B	(.569,.629)B	(.757,.809)B	(.962,.982)B
16	R	(.116,.158)A	(.356,.416)A	(.654,.712)A	(.831,.875)A	(.979,.993)A

NOTE:

[1] S vs. W and W vs. R are the only pairwise comparisons that can be made.

Confidence intervals followed by the same letter are overlapping.

S is the Student pooled t-test.

W is the Welch approximate t-test.

R is the Wilcoxon rank sum test.

Appendix 2. Comparison of Power for $n_1 = n_2 = 12$ under the Normal Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.145,.191)A	(.440,.502)A	(.803,.850)A	(.963,.983)A	(1.00,1.00)A
1	W	(.145,.191)A	(.437,.499)A	(.800,.848)A	(.960,.981)A	(1.00,1.00)A
1	R	(.124,.168)A	(.394,.456)A	(.763,.813)A	(.948,.972)A	(1.00,1.00)A
2	S	(.147,.193)A	(.456,.518)A	(.792,.840)A	(.961,.981)A	(1.00,1.00)A
2	W	(.142,.188)A	(.451,.513)A	(.779,.829)A	(.959,.981)A	(1.00,1.00)A
2	R	(.136,.182)A	(.421,.483)A	(.751,.803)A	(.947,.971)A	(1.00,1.00)A
4	S	(.134,.180)A	(.469,.531)A	(.781,.831)A	(.955,.977)A	(1.00,1.00)A
4	W	(.133,.177)A	(.450,.512)A	(.771,.821)A	(.947,.971)A	(1.00,1.00)A
4	R	(.137,.183)A	(.423,.485)A	(.743,.795)A	(.930,.958)A	(1.00,1.00)A
9	S	(.167,.215)A	(.444,.506)A	(.796,.844)A	(.963,.983)A	(1.00,1.00)A
9	W	(.154,.202)A	(.417,.479)A	(.772,.822)A	(.951,.975)A	(1.00,1.00)A
9	R	-	-	-	-	-
16	S	(.116,.158)A	(.464,.526)A	(.795,.843)A	(.964,.984)A	(1.00,1.00)A
16	W	(.142,.189)A	(.433,.493)A	(.768,.818)A	(.956,.978)A	(1.00,1.00)A
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 3. Comparison of Power for $n_1 = n_2 = 20$ under the Normal Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.122,.166)A	(.452,.514)A	(.812,.858)A	(.958,.980)A	(1.00,1.00)A
1	W	(.122,.166)A	(.450,.512)A	(.810,.856)A	(.957,.980)A	(1.00,1.00)A
1	R	(.130,.174)A	(.427,.490)A	(.793,.841)A	(.952,.976)A	(1.00,1.00)A
2	S	(.149,.195)A	(.456,.518)A	(.809,.855)A	(.962,.982)A	(1.00,1.00)A
2	W	-	-	-	-	-
2	R	(.148,.194)A	(.433,.495)A	(.775,.825)A	(.946,.970)A	(1.00,1.00)A
4	S	(.157,.205)A	(.468,.530)A	(.797,.845)A	(.962,.982)A	(1.00,1.00)A
4	W	(.153,.201)A	(.458,.520)A	(.790,.838)A	(.960,.981)A	(1.00,1.00)A
4	R	(.157,.205)A	(.449,.511)A	(.766,.816)A	(.946,.970)A	(1.00,1.00)A
9	S	(.134,.180)A	(.494,.556)A	(.792,.840)A	(.961,.981)A	(1.00,1.00)A
9	W	(.128,.172)A	(.475,.537)A	(.773,.823)A	(.959,.981)A	(1.00,1.00)A
9	R	-	-	-	-	-
16	S	(.130,.174)A	(.446,.508)A	(.804,.850)A	(.969,.987)A	(1.00,1.00)A
16	W	(.121,.165)A	(.431,.493)A	(.788,.836)A	(.963,.983)A	(1.00,1.00)A
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 4. Comparison of Power for $n_1 = 4$ and $n_2 = 8$ under the Normal Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.121,.165)A	(.454,.516)A	(.739,.791)A	(.934,.962)A	(1.00,1.00)A
1	W	(.114,.156)A	(.402,.464)A	(.658,.716)A	(.880,.918) B	(.984,.996) B
1	R	(.117,.159)A	(.407,.469)A	(.690,.746)A	(.908,.940)AB	(.995,1.00)AB
2	S	–	–	–	–	–
2	W	(.121,.165)	(.400,.417)	(.725,.779)	(.922,.952)	(.997,1.00)
2	R	–	–	–	–	–
4	S	–	–	–	–	–
4	W	–	–	–	–	–
4	R	–	–	–	–	–
9	S	–	–	–	–	–
9	W	(.113,.155)	(.409,.471)	(.730,.784)	(.916,.948)	(.992,.999)
9	R	–	–	–	–	–
16	S	–	–	–	–	–
16	W	(.114,.156)	(.405,.467)	(.719,.773)	(.932,.960)	(1.00,1.00)
16	R	–	–	–	–	–

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 5. Comparison of Power for $n_1 = 8$ and $n_2 = 4$ under the Normal Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.097,.137)A	(.369,.429)A	(.744,.796)A	(.938,.964)A	(1.00,1.00)A
1	W	(.095,.135)A	(.343,.403)A	(.652,.710) B	(.875,.913)B	(.986,.998) B
1	R	(.104,.144)A	(.350,.410)A	(.700,.756)AB	(.918,.949)A	(.995,1.00)AB
2	S	-	-	-	-	-
2	W	(.124,.168)	(.319,.379)	(.608,.668)	(.825,.869)	(.981,.995)
2	R	-	-	-	-	-
4	S	-	-	-	-	-
4	W	(.093,.133)	(.322,.382)	(.576,.636)	(.795,.843)	(.971,.989)
4	R	-	-	-	-	-
9	S	-	-	-	-	-
9	W	(.112.154)	(.310,.368)	(.560,.620)	(.763,.813)	(.997,1.00)
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	-	-	-	-	-
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 6. Comparison of Power for n_1 = 5 and n_2 = 15 under the Normal Distribution.

				δ		
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.120,.164)A	(.456,.518)A	(.765,.815)A	(.959,.981)A	(.997,1.00)A
1	W	(.121,.165)A	(.392,.454) B	(.672,.728)B	(.913,.945) B	(.993,1.00)A
1	R	(.106,.148)A	(.402,.464)AB	(.702,.758)B	(.939,.965)AB	(.997,1.00)A
2	S	-	-	-	-	-
2	W	-	-	-	-	-
2	R	-	-	-	-	-
4	S	-	-	-	-	-
4	W	(.128,.172)	(.429,.491)	(.753,.805)	(.943,.969)	(1.00,1.00)
4	R	-	-	-	-	-
9	S	-	-	-	-	-
9	W	(.119,.163)	(.442,.504)	(.791,.839)	(.942,.968)	(1.00,1.00)
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	-	-	-	-	-
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 7. Comparison of Power for $n_1 = 15$ and $n_2 = 5$ under the Normal Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.118,.162)A	(.446,.508)A	(.771,.821)A	(.936,.964)A	(1.00,1.00)A
1	W	(.122,.166)A	(.384,.446)A	(.693,.749)B	(.888,.924) B	(.997,1.00)A
1	R	(.106,.148)A	(.398,.460)A	(.714,.768)B	(.914,.946)AB	(1.00,1.00)A
2	S	-	-	-	-	-
2	W	(.107,.149)A	(.387,.449)B	(.666,.724)B	(.849,.891)B	(.991,.999)A
2	R	(.121,.165)A	(.463,.525)A	(.770,.820)A	(.928,.956)A	(.997,1.00)A
4	S	-	-	-	-	-
4	W	(.127,.171)	(.357,.417)	(.599,.659)	(.853,.895)	(.985,.997)
4	R	-	-	-	-	-
9	S	-	-	-	-	-
9	W	(.115,.157)	(.303,.361)	(.626,.684)	(.838,.882)	(.994,1.00)
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	(.099,.139)	(.315,.373)	(.608,.668)	(.824,.868)	(.984,.996)
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 8. Comparison of Power for $n_1 = 12$ and $n_2 = 20$ under the Normal Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.142,.188)A	(.460,.522)A	(.787,.835)A	(.958,.980)A	(1.00,1.00)A
1	W	(.139,.185)A	(.453,.515)A	(.777,.827)A	(.951,.975)A	(1.00,1.00)A
1	R	(.133,.177)A	(.436,.498)A	(.765,.815)A	(.943,.969)A	(1.00,1.00)A
2	S	(.096,.136)A[1]	(.394,.456)A[1]	(.742,.794)A[1]	(.950,.974)A[1]	(.997,1.00)A[1]
2	W	(.120,.164)	(.461,.523)	(.798,.846)	(.970,.988)	(1.00,1.00)
2	R	(.108,.150)A	(.403,.465)A	(.734,.786)A	(.947,.971)A	(.997,1.00)A
4	S	-	-	-	-	-
4	W	(.155,.203)A	(.459,.521)A	(.810,.856)A	(.955,.977)A	(1.00,1.00)A
4	R	(.113,.155)A	(.377,.434)B	(.712,.766)B	(.928,.956)B	(1.00,1.00)B
9	S	-	-	-	-	-
9	W	(.139,.185)	(.455,.517)	(.807,.853)	(.959,.981)	(1.00,1.00)
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	(.132,.176)A	(.441,.503)A	(.798,.846)A	(.970,.988)A	(1.00,1.00)A
16	R	(.089,.127)B	(.341,.401)B	(.676,.732)B	(.886,.922)B	(.995,1.00)A

NOTE:

[1] S vs. R is the only pairwise comparison that can be made.
Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 9. Comparison of Power for n_1 = 20 and n_2 = 12 under the Normal Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.138,.184)A	(.434,.496)A	(.799,.847)A	(.962,.982)A	(1.00,1.00)A
1	W	(.132,.176)A	(.410,.472)A	(.789,.837)A	(.959,.981)A	(1.00,1.00)A
1	R	(.124,.168)A	(.402,.464)A	(.775,.825)A	(.952,.976)A	(1.00,1.00)A
2	S	-	-	-	-	-
2	W	(.110,.152)A	(.413,.475)A	(.789,.837)A	(.956,.978)A	(1.00,1.00)A
2	R	(.126,.170)A	(.435,.497)A	(.800,.848)A	(.950,.974)A	(1.00,1.00)A
4	S	-	-	-	-	-
4	W	(.112,.154)	(.440,.502)	(.775,.825)	(.950,.974)	(1.00,1.00)
4	R	-	-	-	-	-
9	S	-	-	-	-	-
9	W	(.142,.188)	(.439,.501)	(.777,.827)	(.947,.971)	(1.00,1.00)
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	(.145,.191)	(.432,.494)	(.794,.842)	(.939,.965)	(1.00,1.00)
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 10. Comparison of Power for $n_1 = n_2 = 4$ under the Double Exponential Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.129,.173)	(.431,.493)	(.713,.767)	(.874,.912)	(.984,.996)
1	W	-	-	-	-	-
1	R	-	-	-	-	-
2	S	(.148,.194)A	(.448,.510)A	(.729,.783)A	(.876,.914)A	(.983,.995)A
2	W	(.118,.160)A	(.391,.453)A	(.674,.730)A	(.833,.877)A	(.971,.989)A
2	R	-	-	-	-	-
4	S	(.125,.169)	(.470,.532)	(.725,.779)	(.848,.890)	(.973,.989)
4	W	-	-	-	-	-
4	R	-	-	-	-	-
9	S	(.177,.227)A	(.486,.548)A	(.716,.770)A	(.879,.917)A	(.962,.982)A
9	W	-	-	-	-	-
9	R	(.157,.205)A	(.426,.488)A	(.658,.716)A	(.813,.859)B	(.938,.964)A
16	S	(.166,.214)A	(.482,.544)A	(.716,.770)A	(.874,.912)A	(.967,.985)A
16	W	-	-	-	-	-
16	R	(.154,.202)A	(.452,.514)A	(.676,.732)A	(.809,.855)B	(.947,.971)A

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 11. Comparison of Power for $n_1 = n_2 = 12$ under the Double Exponential Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.145,.191)A	(.465,.527)AB	(.813,.859)A	(.950,.974)A	(.995,1.00)A
1	W	(.139,.185)A	(.459,.521) B	(.807,.853)A	(.947,.971)A	(.995,1.00)A
1	R	(.152,.200)A	(.524,.586)A	(.848,.890)A	(.962,.982)A	(.997,1.00)A
2	S	(.141,.187)A	(.505,.567)A	(.790,.838)B	(.940,.966)A	(1.00,1.00)A
2	W	(.137,.183)A	(.492,.554)A	(.784,.832)B	(.936,.964)A	(1.00,1.00)A
2	R	(.168,.216)A	(.542,.604)A	(.834,.885)A	(.995,.977)A	(1.00,1.00)A
4	S	(.149,.195)AB	(.490,.552)B	(.804,.850)AB	(.932,.960)A	(.995,1.00)A
4	W	(.140,.186) B	(.478,.540)B	(.786,.834) B	(.930,.958)A	(.995,1.00)A
4	R	(.195,.247)A	(.569,.629)A	(.846,.888)A	(.950,.974)A	(1.00,1.00)A
9	S	(.155,.201)A	(.489,.551)A	(.812,.858)A	(.935,.963)A	(.994,1.00)A
9	W	(.149,.195)A	(.463,.525)A	(.790,.838)A	(.925,.955)A	(.994,1.00)A
9	R	-	-	-	-	-
16	S	(.172,.222)BA[1]	(.505,.567)BA[1]	(.807,.853)BA[1]	(.926,.956)BA[1]	(.994,1.00)AA[1]
16	W	(.153,.201) A	(.473,.535) A	(.789,.837) A	(.914,.946) A	(.994,1.00) A
16	R	(.234,.288)A	(.620,.680)A	(.867,.907)A	(.968,.986)A	(.997,1.00)A

NOTE:

[1] S vs. R and S vs. W are the only pairwise comparisons that can be made.
Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 12. Comparison of Power for $n_1 = n_2 = 20$ under the Double Exponential Distribution.

				δ		
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.143,.189)B	(.487,.549)B	(.792,.840)B	(.951,.975)B	(1.00,1.00)A
1	W	(.142,.188)B	(.487,.549)B	(.791,.840)B	(.951,.975)B	(1.00,1.00)A
1	R	(.203,.255)A	(.580,.640)A	(.875,.913)A	(.980,.994)A	(1.00,1.00)A
2	S	(.134,.178)B	(.471,.533)B	(.792,.840)B	(.962,.982)A	(1.00,1.00)A
2	W	(.131,.175)B	(.465,.527)B	(.788,.836)B	(.959,.981)A	(1.00,1.00)A
2	R	(.181,.231)A	(.592,.652)A	(.868,.908)A	(.974,.990)A	(1.00,1.00)A
4	S	(.153,.201)B	(.482,.544)B	(.798,.846)B	(.957,.979)B	(.997,1.00)A
4	W	(.148,.194)B	(.475,.537)B	(.794,.842)B	(.955,.977)B	(.997,1.00)A
4	R	(.206,.258)A	(.604,.664)A	(.891,.927)A	(.985,.997)A	(1.00,1.00)A
9	S	(.163,.211)A	(.525,.587)A	(.793,.841)A	(.954,.976)A	(1.00,1.00)A
9	W	(.148,.194)A	(.510,.571)A	(.784,.832)A	(.951,.975)A	(.997,1.00)A
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	-	-	-	-	-
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 13. Comparison of Power for $n_1 = 4$ and $n_2 = 8$ under the Double Exponential Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	-	-	-	-	-
1	W	-	-	-	-	-
1	R	(.161,.209)	(.469,.531)	(.712,.766)	(.879,.917)	(.979,.993)
2	S	-	-	-	-	-
2	W	(.105,.147)A	(.439,.501)A	(.737,.789)A	(.865,.905)A	(.983,.995)A
2	R	(.118,.160)A	(.411,.473)A	(.725,.779)A	(.871,.909)A	(.985,.997)A
4	S	-	-	-	-	-
4	W	(.116,.158)A	(.432,.494)A	(.742,.794)A	(.909,.941)A	(.986,.998)A
4	R	(.123,.167)A	(.384,.446)A	(.678,.734)B	(.876,.914)A	(.977,.993)A
9	S	-	-	-	-	-
9	W	(.151,.199)	(.450,.512)	(.736,.789)	(.889,.925)	(.991,.999)
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	(.143,.189)	(.466,.528)	(.728,.782)	(.879,.917)	(.905,.939)
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 14. Comparison of Power for $n_1 = 8$ and $n_2 = 4$ under the Double Exponential Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	-	-	-	-	-
1	W	-	-	-	-	-
1	R	(.149,.195)	(.471,.533)	(.710,.764)	(.882,.920)	(.984,.996)
2	S	-	-	-	-	-
2	W	(.150,.196)	(.420,.482)	(.680,.736)	(.836,.880)	(.954,.976)
2	R	-	-	-	-	-
4	S	-	-	-	-	-
4	W	(.148,.194)	(.412,.474)	(.648,.706)	(.793,.841)	(.936,.964)
4	R	-	-	-	-	-
9	S	-	-	-	-	-
9	W	-	-	-	-	-
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	-	-	-	-	-
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 15. Comparison of Power for n_1 = 5 and n_2 = 15 under the Double Exponential Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.159,.207)A	(.457,.519)A	(.787,.835)A	(.942,.968)A	(1.00,1.00)A
1	W	(.146,.192)A	(.425,.487)A	(.736,.788)A	(.893,.929) B	(.980,.994)B
1	R	(.162,.210)A	(.453,.515)A	(.776,.826)A	(.925,.955)AB	(.985,.997)B
2	S	-	-	-	-	-
2	W	(.145,.191)	(.439,.501)	(.770,.820)	(.898,.932)	(.989,.999)
2	R	-	-	-	-	-
4	S	-	-	-	-	-
4	W	-	-	-	-	-
4	R	-	-	-	-	-
9	S	-	-	-	-	-
9	W	(.147,.193)	(.459,.521)	(.783,.831)	(.939,.965)	(1.00,1.00)
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	(.126,.170)	(.488,.550)	(.781,.831)	(.924,.954)	(.997,1.00)
16	R	-	-	-	-	-

NOTE:

Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 16. Comparison of Power for $n_1 = 15$ and $n_2 = 5$ under the Double Exponential Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.157,.205)A	(.465,.527)A	(.817,.863)A	(.939,.965)A	(1.00,1.00)A
1	W	(.152,.200)A	(.428,.490)A	(.739,.791)B	(.873,.911) B	(.985,.997)B
1	R	(.169,.219)A	(.468,.530A	(.807,.853)A	(.911,.943)AB	(.988,.998)B
2	S	-	-	-	-	-
2	W	(.144,.190)[1]	(.437,.499)[1]	(.698,.754)[1]	(.850,.892)[1]	(.968,.986)[1]
2	R	(.195,.247)	(.526,.588)	(.783,.831)	(.910,.942)	(.977,.993)
4	S	-	-	-	-	-
4	W	-	-	-	-	-
4	R	-	-	-	-	-
9	S	-	-	-	-	-
9	W	-	-	-	-	-
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	-	-	-	-	-
16	R	-	-	-	-	-

NOTE:

[1] No comparison between W and R can be made.
Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 17. Comparison of Power for $n_1 = 12$ and $n_2 = 20$ under the Double Exponential Distribution.

		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.144,.190)A	(.487,.549)B	(.796,.844)B	(.961,.981)AB	(1.00,1.00)A
1	W	(.138,.184)A	(.480,.542)B	(.788,.836)B	(.947,.971) B	(1.00,1.00)A
1	R	(.182,.232)A	(.573,.633)A	(.589,.899)A	(.980,.994)A	(1.00,1.00)A
2	S	-	-	-	-	-
2	W	(.161,.209)A	(.447,.509)B	(.805,.851)A	(.949,.973)A	(1.00,1.00)A
2	R	(.179,.229)A	(.519,.581)A	(.845,.887)A	(.968,.986)A	(.997,1.00)A
4	S	-	-	-	-	-
4	W	(.148,.194)A	(.483,.545)A	(.801,.849)A	(.938,.964)A	(1.00,1.00)A
4	R	(.151,.198)A	(.542,.604)A	(.845,.887)A	(.952,.976)A	(1.00,1.00)A
9	S	-	-	-	-	-
9	W	(.134,.180)A	(.460,.522)B	(.797,.845)B	(.928,.957)A	(1.00,1.00)A
9	R	(.169,.219)A	(.548,.610)A	(.856,.896)A	(.951,.975)A	(.995,1.00)A
16	S	-	-	-	-	-
16	W	-	-	-	-	-
16	R	-	-	-	-	-

NOTE:

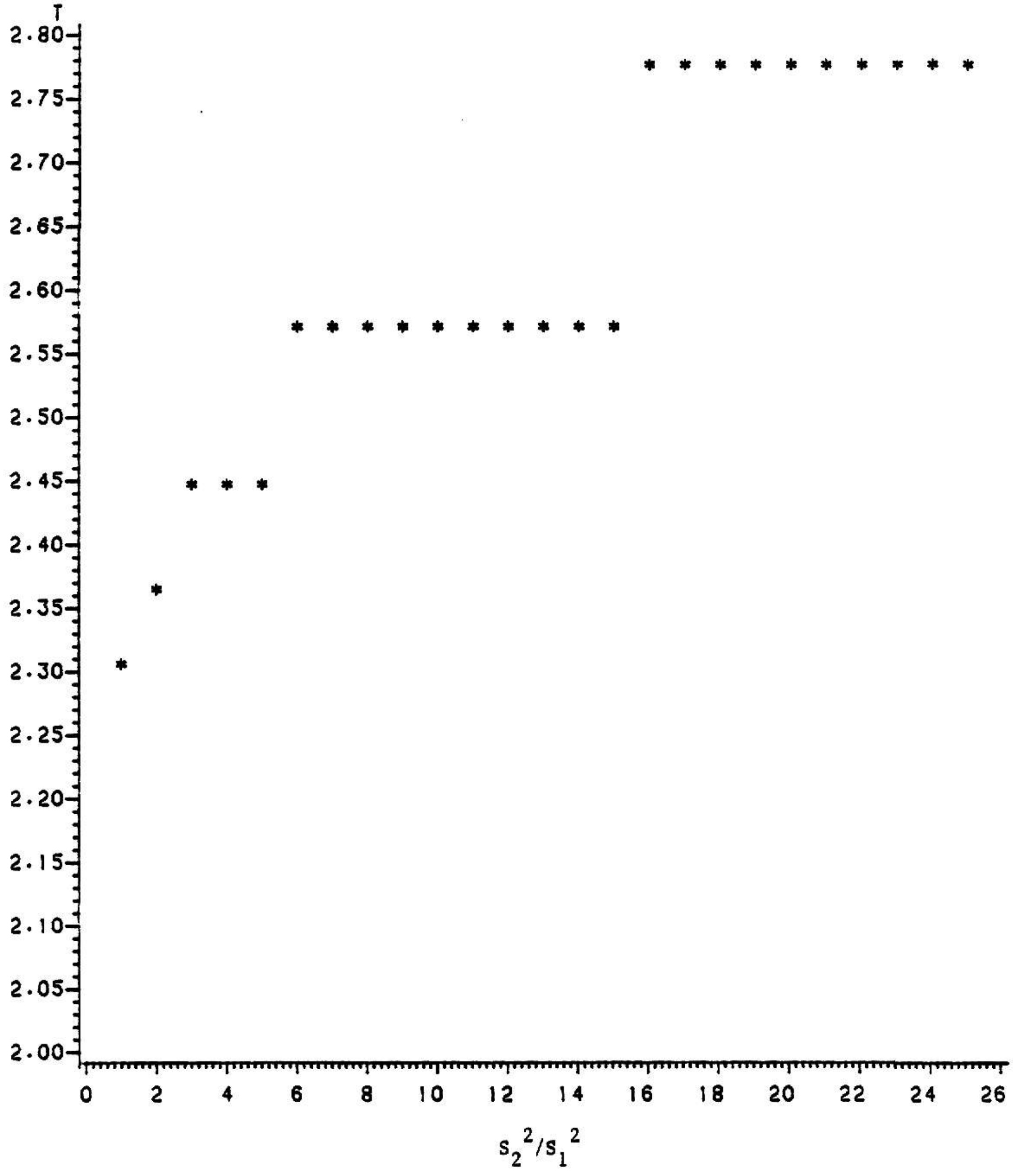
Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.

Appendix 18. Comparison of Power for $n_1 = 20$ and $n_2 = 12$ under the Double Exponential Distribution.

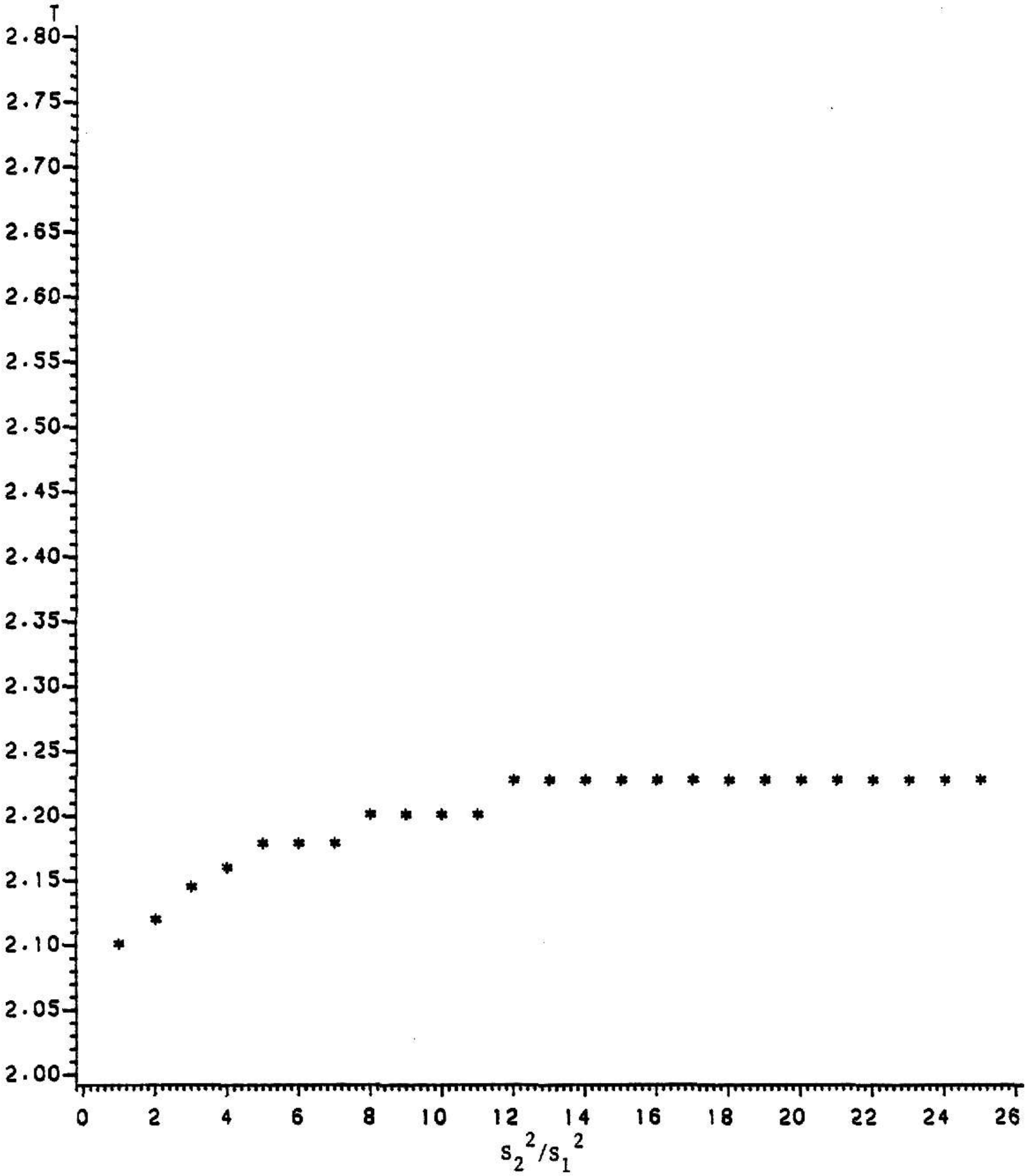
		δ				
σ_2^2/σ_1^2	Test	1	2	3	4	6
1	S	(.135,.181)A	(.489,.551)B	(.785,.833)B	(.948,.972)A	(.997,1.00)A
1	W	-	-	-	-	-
1	R	(.173,.223)A	(.576,.636)A	(.842,.884)A	(.968,.986)A	(.997,1.00)A
2	S	-	-	-	-	-
2	W	(.141,.187)B	(.490,.552)B	(.792,.804)B	(.930,.958)B	(.994,1.00)A
2	R	(.214,.266)A	(.606,.666)A	(.887,.923)A	(.968,.986)A	(1.00,1.00)A
4	S	-	-	-	-	-
4	W	(.152,.200)	(.493,.555)	(.789,.837)	(.942,.968)	(.995,1.00)
4	R	-	-	-	-	-
9	S	-	-	-	-	-
9	W	(.148,.194)	(.472,.534)	(.774,.824)	(.921,.951)	(.988,.998)
9	R	-	-	-	-	-
16	S	-	-	-	-	-
16	W	(.137,.183)	(.469,.531)	(.765,.815)	(.920,.950)	(.989,.999)
16	R	-	-	-	-	-

NOTE:

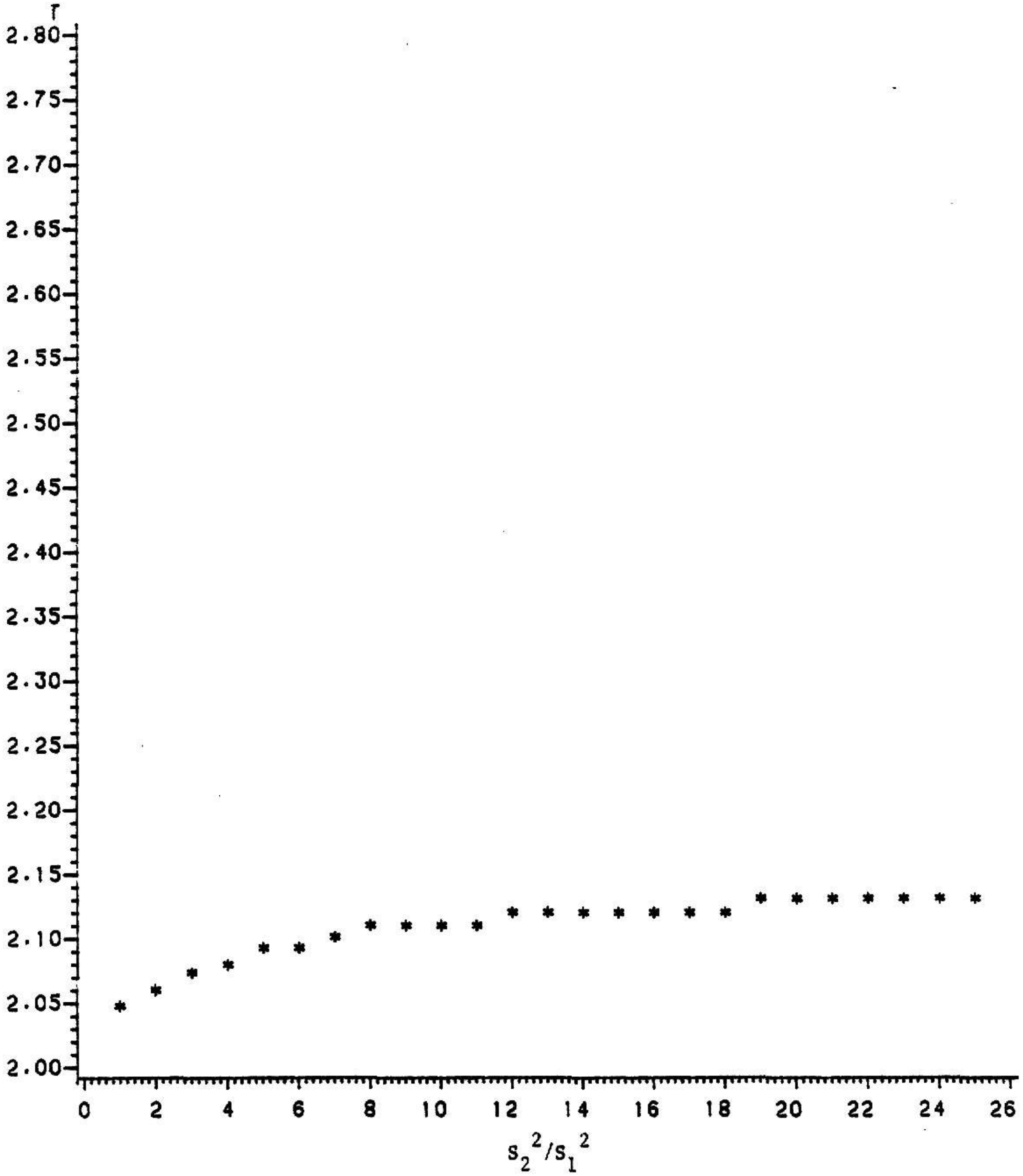
Confidence intervals followed by the same letter are overlapping.
S is the Student pooled t-test.
W is the Welch approximate t-test.
R is the Wilcoxon rank sum test.



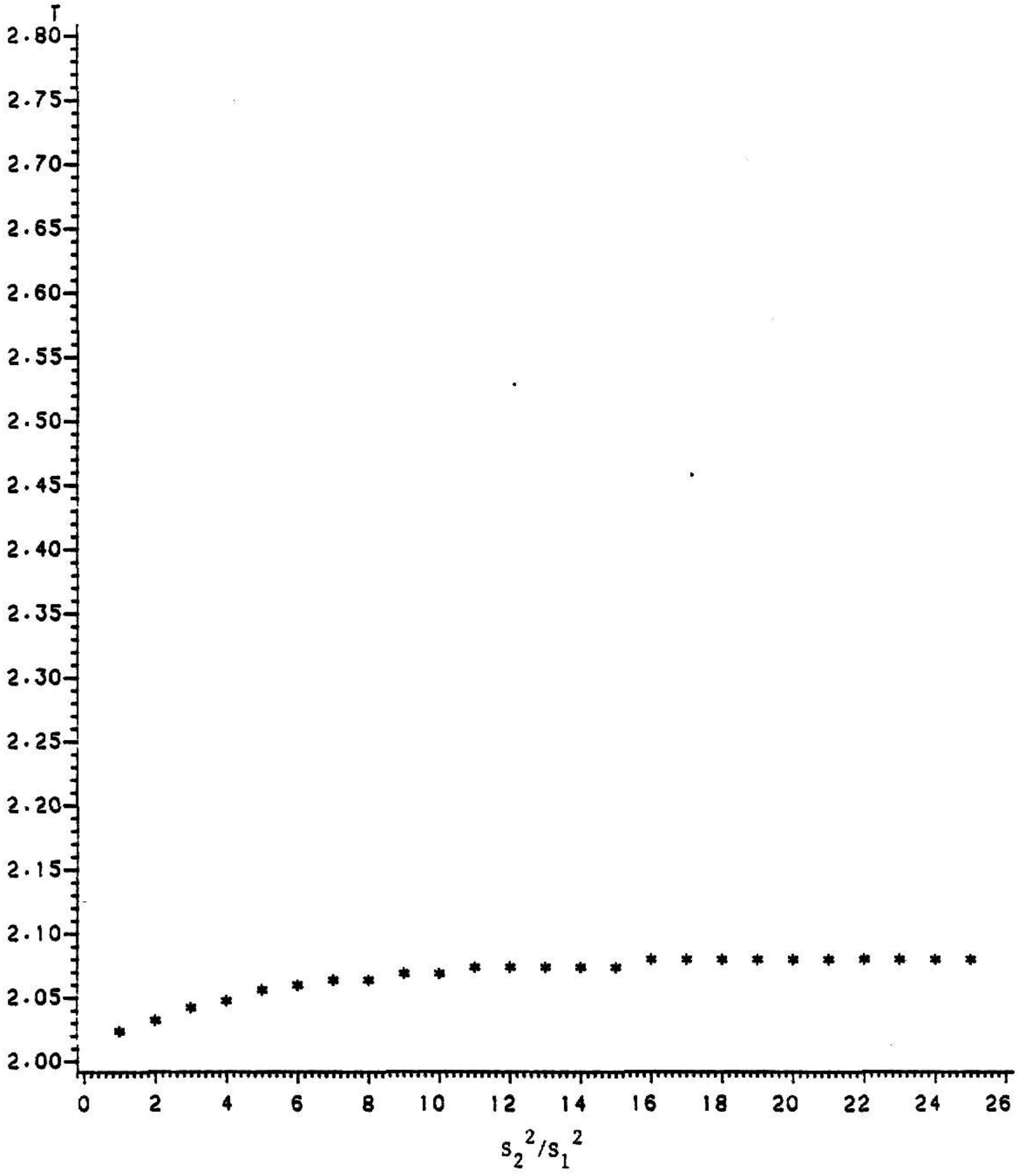
Appendix 19. Relationship Between S_2^2/S_1^2 and the Critical Values of Welch's Approximate t-Test for $n_1=n_2=5$ where $t_{.025}^{(8)}=2.306$.



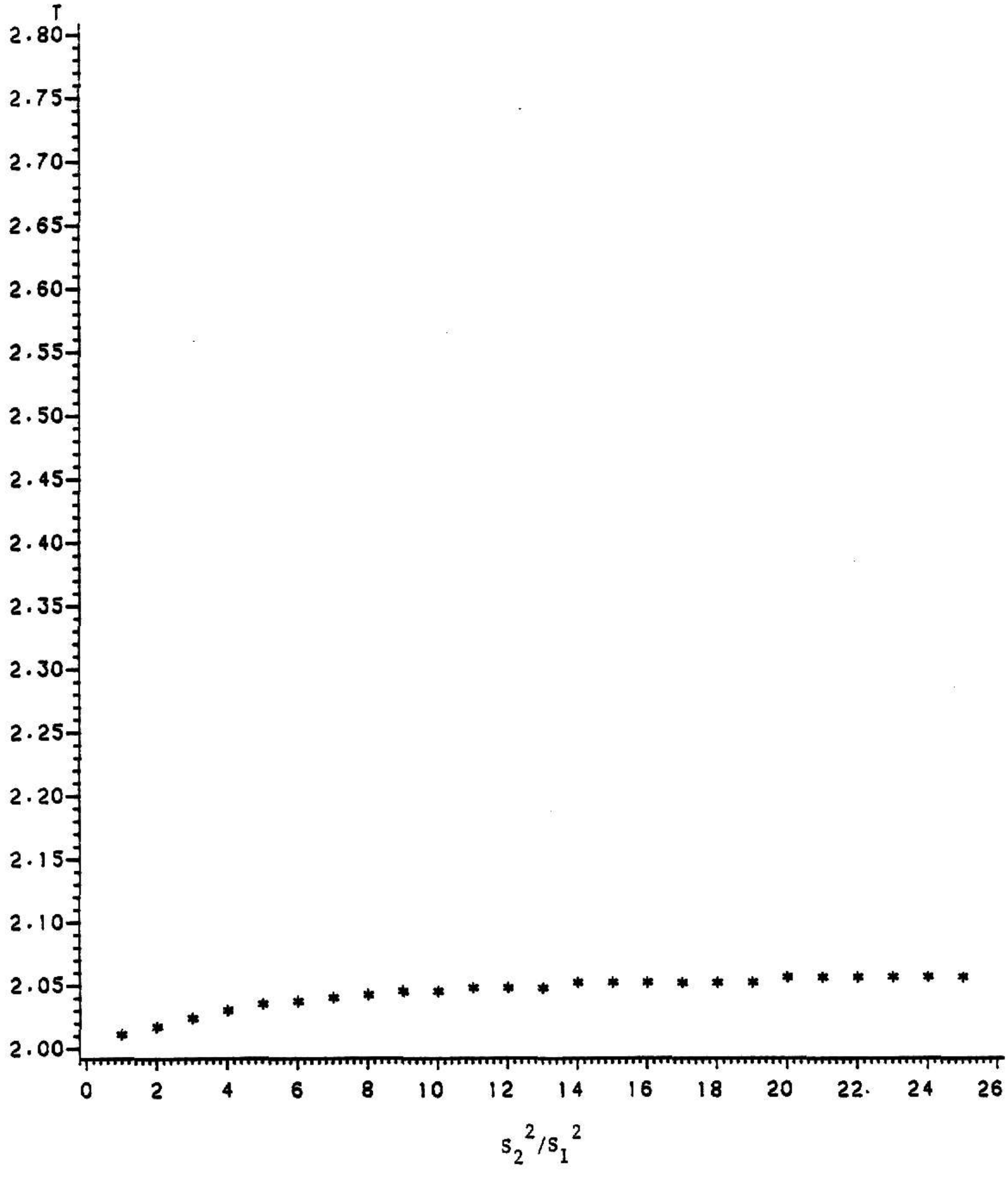
Appendix 20. Relationship Between S_2^2/S_1^2 and the Critical Values of Welch's Approximate t-Test for $n_1=n_2=10$ where $t_{.025}^{(18)}=2.101$.



Appendix 21. Relationship Between S_2^2/S_1^2 and the Critical Values of Welch's Approximate t-Test for $n_1=n_2=15$ where $t_{.025}^{(28)}=2.048$.



Appendix 22. Relationship Between S_2^2/S_1^2 and the Critical Values of Welch's Approximate t-Test for $n_1=n_2=20$ where $t_{.025}^{(38)}=2.024$.



Appendix 23. Relationship Between S_2^2/S_1^2 and the Critical Values of Welch's Approximate t-Test for $n_1=n_2=25$ where $t_{.025}^{(48)}=2.011$

A COMPARISON OF THE PERFORMANCE OF SEVERAL SOLUTIONS TO THE BEHRENS-FISHER PROBLEM

by

BARBARA ROSE KUZMAK

B. S., Cornell University, 1978
M. S., Kansas State University, 1982

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1986

ABSTRACT

I evaluate the performance of Student's pooled t-test, Welch's approximate t-test and Wilcoxon's rank sum test under several combinations of variances, sample sizes and population distributions. The study investigates 540 combinations by computer simulation. I appraise the tests in terms of significance level and power. Test recommendation is dependent on sample size and population distribution. Equality or inequality of sample size determines whether or not equality of variances is important. If sample sizes are equal, then tests which perform well under equal variances are insensitive to departures of variance. This is true for the performance of Student's test under either distribution. Likewise, Wilcoxon's test displays a similar behavior for equal sample sizes under the double exponential distribution. In contrast, these tests perform poorly under both distributions when sample sizes and variances are unequal. For these situations, I recommend Welch's test.