Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 14

11-1-2003

A Nonparametric Fitted Test For The Behrens-Fisher Problem

Terry Hyslop
Thomas Jefferson University, terry.hyslop@jefferson.edu

Paul J. Lupinacci
Villanova University, Paul.Lupinacci@villanova.edu

DOI: 10.22237/jmasm/1067645640

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 414-424

1538 – 9472/03/$30.00

A Nonparametric Fitted Test For The Behrens-Fisher Problem

Terry Hyslop
Department of Medicine
Thomas Jefferson University

Paul J. Lupinacci
Department of Mathematical Sciences
Villanova University

A nonparametric test for the Behrens-Fisher problem that is an extension of a test proposed by Fligner and Policello was developed. Empirical level and power estimates of this test are compared to those of alternative nonparametric and parametric tests through simulations. The results of our test were better than or comparable to all tests considered.

Key words: Behrens-Fisher problem, empirical level and power, Wilcoxon-Mann-Whitney, nonparametric, simulation study

Introduction

The comparison of the means of two independent populations has traditionally been approached using Student's t-test. The use of this test assumes that the observations come from a normal distribution and that the variances of the two populations are equal. When the homogeneity of variances is not a reasonable assumption the problem has been called the Behrens-Fisher problem.

Lee and Gurland (1975) developed a new method for handling the Behrens-Fisher problem and compared their test to many others that have been proposed for this problem. Their test performed very well regarding size and power. However, their method utilized a large table of critical values to determine the correct region of rejection. Lee and Fineberg (1991) sought to simplify the method proposed by Lee and Gurland. They fit a nonlinear function to the critical values derived by Lee and Gurland so that the critical values could be estimated.

Terry Hyslop is Assistant Professor of Medicine in the Biostatistics Section, Division of Clinical Pharmacology, Department of Medicine at Thomas Jefferson University. Email: Terry.Hyslop@jefferson.edu. Paul J. Lupinacci is Assistant Professor in the Department of Mathematical Sciences at Villanova University. E-mail: Paul.Lupinacci@villanova.edu.

Various authors have also considered the Behrens-Fisher problem when the normality assumption is not appropriate. The usual nonparametric approaches assume that the data are continuous and the distributions are of the same shape. For these tests, such as the Wilcoxon-Mann-Whitney test (Wilcoxon 1945; Mann & Whitney 1947), the level of the test will not be preserved when the populations have different shapes or variances (Fligner & Policello 1981; Brunner & Neumann 1982, 1986; Brunner & Munzel 2000). Fligner and Policello (1981) and Brunner and Neumann (1982, 1986) considered the problem under the assumption that the independent samples are from continuous distributions without the assumption of equal variances or equal shapes of the distributions. Brunner and Munzel (2000) derived an asymptotically distribution free test without the assumption that the data are generated from a continuous distribution function.

Fligner and Policello developed their alternative nonparametric method for comparing two population medians without the equal variance and equal shape assumptions. To implement their test, one must consult a large table of critical values to determine the correct region of rejection. Their table is parameterized by the test's level of significance and the sample sizes of the two samples. We expand on the approach of Fligner and Policello by proposing a fitted test which eliminates the need for large tables or complicated derivations of critical values. We fit a nonlinear function to the critical

values in their table so that the critical values can be estimated. Motivation for the nonlinear function came from the nonlinear function used by Lee and Fineberg. A complete description of the problem and details of the proposed test follow in the Methodology Section. In that section, our method is demonstrated using a numerical example. Simulation studies are used in the Results Section to compare the fitted test to some of the other parametric and nonparametric tests which have been proposed for the Behrens-Fisher problem.

Methodology

Let $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ be independent random samples from continuous distributions with population medians θ_x and θ_y, respectively. We are interested in testing the following hypotheses:

$$H_0:\theta_x=\theta_y$$
$$\text{versus } H_a:\theta_x>\theta_y\left[\text{or } \theta_x<\theta_y \text{ or } \theta_x\neq\theta_y\right].$$

Let P_i represent the number of sample observations, Y_j, less than X_i, for $i=1,\ldots,m$. Similarly, let Q_j represent the number of sample observations, X_i, less than Y_j, for $j=1,\ldots,n$. Compute the average placement for each of the samples,

$$\bar{P}=\frac{1}{m}\sum_{i=1}^{m}P_i \text{ and } \bar{Q}=\frac{1}{n}\sum_{j=1}^{n}Q_j .$$

$$\text{Let } V_1=\sum_{i=1}^{m}\left(P_i-\bar{P}\right)^2 \text{ and } V_2=\sum_{j=1}^{n}\left(Q_j-\bar{Q}\right)^2,$$

and calculate the test statistic

$$\hat{U}=\frac{\sum_{i=1}^{m}P_i-\sum_{j=1}^{n}Q_j}{2\left(V_1+V_2+\bar{P}\,\bar{Q}\right)^{1/2}} .$$

Fligner and Policello presented a test of H_0 based on this statistic, where the procedure at an approximate α level of significance versus the one-sided alternative $\theta_x>\theta_y$ is

Reject H_0 if $\hat{U}\geq u_\alpha$; otherwise do not reject.

They provided a table of critical values for u_α for various values of m, n, and α. Values outside the range of their table are to be derived or estimated by z_α for large sample sizes, where z_α is the 1-α percentile of the standard normal distribution.

Implementation of this test would be greatly simplified if the large table of critical values was not required. In addition, sample size combinations that are not provided in their table would require either additional effort for derivation, or an assumption of $u_\alpha=z_\alpha$. We propose fitting the following function to the critical values in the Fligner and Policello table so that the critical values can be estimated:

$$u_\alpha=b_0+b_1/f_1+b_2/f_2+b_3/(f_1f_2)+b_4/f_1^2+b_5/f_2^2,$$

where $f_1=m-1$, $f_2=n-1$, and $b_{0,\ldots,}b_5$ are the parameters of the function. We also propose that the parameters $b_0,\ldots,b_5$ be estimated by ordinary least squares. 54 values obtained from Fligner and Policello's table of critical values were used in the estimation process. Table 1 presents the parameter estimates obtained for the various α values of 0.10, 0.05, 0.025, and 0.01.

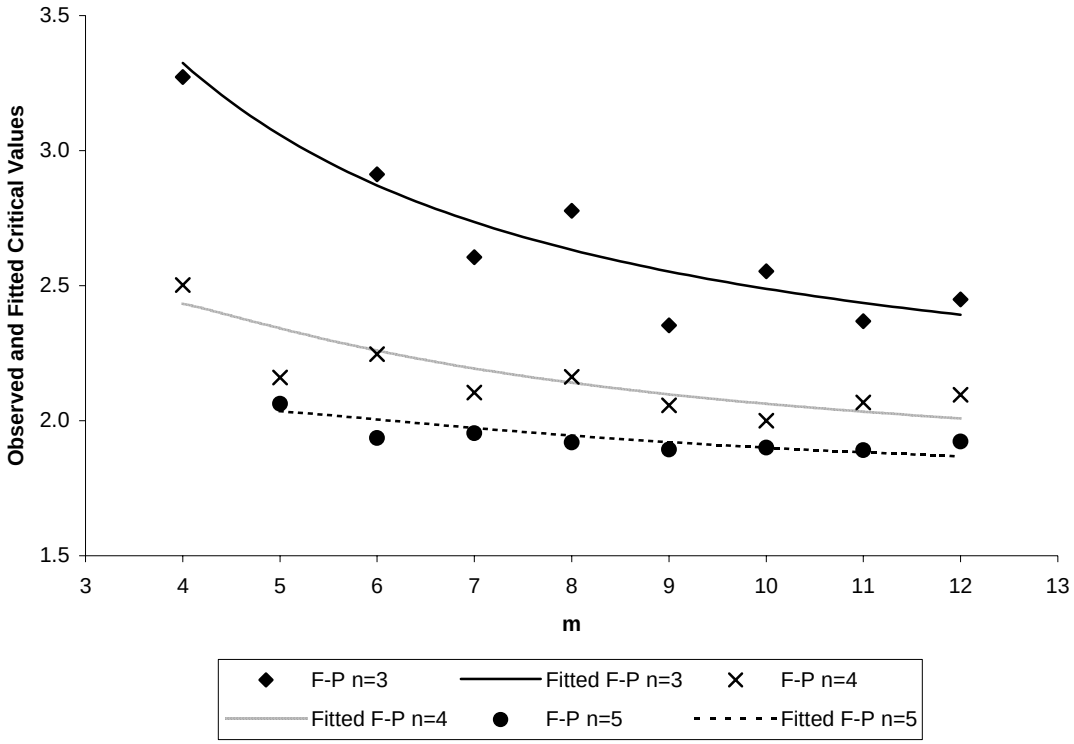
Table 1. Parameter estimates for the F-P fitted test polynomial.

α	b_0	b_1	b_2	b_3	b_4	b_5
0.10	1.34	-1.39	0.16	-0.03	5.20	1.17
0.05	1.74	-0.69	-0.87	12.53	-4.09	2.44
0.025	2.15	-0.60	-2.54	22.05	-3.50	7.36
0.01	3.16	-11.43	-6.75	50.15	51.87	19.20

Motivation for this functional form comes from a parametric fitted test for the Behrens-Fisher problem proposed by Lee and Fineberg (1991) as an alternative to Lee and Gurland's (1975) test that also required

extensive tables of critical values. Their proposed function is similar to the one proposed here. Other functional forms were also considered, but none were found which provided a better fit to the critical values. Figure 1 displays the critical and fitted values for a level 0.05 test when m = 4, 5, …, 12 and n = 3, 4, and 5. The fit is good for even these small values of n and becomes more precise as n gets larger. The test based on the fitted critical values will be referred to as $\hat{U}_f$, and the critical value will be referred to as $u_\alpha^{(f)}$. For example, Fligner and Policello's critical value for a one-sided, level 0.05 test when both samples are of size 5 is 2.063. Using our parameter estimates when $\alpha = 0.05$ and the sample sizes, we obtain an estimated critical value of 2.035.

Figure 1. Plot of Fligner and Policello's Critical Values and the Fitted Critical Values for m=4(1)12 and n=3,4,5.



Numerical Example

The following example uses a data set that originated from the simulation studies that are presented in the Results Section of the manuscript. With this data set, we will test the null hypothesis that the two population medians are the same versus the alternative hypothesis that the median of the first population is greater than that of the second population, that is,

$$H_o : \theta_x = \theta_y \text{ versus } H_a : \theta_x > \theta_y .$$

The data in both groups were simulated from uniform distributions with a mean of 100. However, the variance of the second distribution was ten times that of the first distribution. The first data set consists of twelve observations while the second data set consists of only five observations. Thus, we are simulating a scenario where the data set with fewer observations

comes from an underlying distribution function with a larger variance. We will utilize this data set to demonstrate the test procedure and to illustrate the need for a simulation study which compares the various methods that are used for analyzing this type of data in terms of their power and ability to hold the level of significance. We will use the notation as defined in the Methodology Section. The data, as well as the placement values, P_i and Q_j, for each observation in the first and second samples, respectively, are given in Table 2.

Table 2. Data for the Numerical Example

Group 1			Group 2		
Observation	Value	P_i	Observation	Value	Q_j
1	101.673	4	1	103.409	12
2	101.550	4	2	98.546	1
3	100.410	4	3	97.429	0
4	100.203	4	4	96.536	0
5	99.906	4	5	95.940	0
6	99.875	4			
7	99.861	4			
8	99.695	4			
9	99.535	4			
10	98.985	4			
11	98.575	4			
12	98.461	3			
$\sum_{i=1}^{12} P_i = 47$			$\sum_{i=1}^{5} Q_j = 13$		

For this example, the sum of the placements in the first data set is 47 and the sum of the placements in the second data set is 13. this leads to the average placements of:

$$\bar{P} = \frac{1}{12}\sum_{i=1}^{12} P_i = 3.917$$

and

$$\bar{Q} = \frac{1}{5}\sum_{j=1}^{5} Q_j = 2.600$$

for each group. The values of V_1 and V_2 are

$$V_1 = \sum_{i=1}^{12}\left(P_i - \bar{P}\right)^2 = 0.917$$

and

$$V_2 = \sum_{j=1}^{5}\left(Q_j - \bar{Q}\right)^2 = 111.200,$$

and the test statistic is calculated as

$$\hat{U} = \frac{\sum_{i=1}^{12} P_i - \sum_{j=1}^{5} Q_j}{2\left(V_1 + V_2 + \bar{P}\,\bar{Q}\right)^{1/2}} = 1.537.$$

For m = 12 and n = 5, the critical value for the fitted test is $u_{0.05}^{(f)}$ = 1.868, and the critical value for the Fligner and Policello test is $u_{0.05}$ = 1.923. Therefore, we fail to reject the null hypothesis using both tests. However, the calculation of the Fligner and Policello critical value would have been much more complicated if our sample sizes were not given in their table of critical values. Therefore, we suggest using the critical value based on the fitted test.

Let us also consider how alternative

tests for this type of data would have fared with this data set. Since the data are not coming from a normal distribution, most statisticians would use the nonparametric alternative to Student's t test, the Wilcoxon-Mann-Whitney test, without hesitation. For this example, we compared our results to those of the Wilcoxon-Mann-Whitney test and the nonparametric test proposed by Brunner and Munzel (2000). The Brunner and Munzel test led to the same conclusion as our fitted test, that is, their test failed to reject the null hypothesis. The Brunner and Munzel test statistic was $B = 1.437$ and the corresponding p-value was 0.112, which is not significant at the 0.05 level of significance. However, there was a different result if one used the Wilcoxon-Mann-Whitney test. Its test statistic was $W = 28$ and the corresponding p-value was 0.041, which is significant at the 0.05 level of significance. This conflicting result caught our attention and spurred interest in a simulation study which compares the various methods in terms of size and power.

Results

For our simulation study, we considered three nonparametric procedures and one parametric procedure. The three nonparametric procedures that were considered were the Wilcoxon-Mann-Whitney test, denoted W, the Brunner and Munzel test, denoted B, and our fitted test, denoted $\hat{U}_f$. The parametric test that we included in our simulation study was the usual t test using Satterthwaite's approximation for the degrees of freedom. We used t_s to denote this test. We decided not to include the Fligner and Policello test in the discussion because its empirical level and power estimates were almost identical to those of our fitted test. This was to be expected since we fitted a function to their critical values and the fit was very good. We simulated data using the normal, contaminated normal, double exponential, uniform, and gamma distributions for estimating both the empirical level and power for the four tests. Since we are interested in determining the effect of different variances on the level and power estimates, we considered distributions which differed in scale by assuming that if

$$X_1,\ldots,X_m \sim F(x),$$

then we let

$$Y_1,\ldots,Y_n \sim G(y) = F\left(y/\sigma\right)$$

for values of

$$\sigma^2 = \{0.01, 0.25, 1, 4, 10\}.$$

All simulations were run in SAS version 8. The SAS function NORMAL was used to generate random standard normal deviates which were then transformed to simulate the desired normal distribution. The contaminated normal deviates were generated by multiplying a random normal deviate by 9 with probability p = 0.10. The double exponential deviates were generated using the method of Martinez and Iglewicz (1984) that transforms a random standard normal deviate into a double exponential deviate using the transformation

$$DE = Z^{\exp\{\frac{0.109Z^2}{2}\}},$$

where Z is a random standard normal deviate. Random uniform and gamma deviates were generated using the SAS functions UNIFORM and RANGAM, respectively.

For a statistical test to be meaningful, it must display adequate power while still maintaining its nominal level. We ran simulations to obtain estimates of the level and power for each of the tests under consideration. To estimate the tests' level, we ran 15,500 simulation iterations. The number of simulations provides that a 95% confidence interval for the estimated level will be approximately $\pm 0.36\%$ for $\alpha = 0.05$. At each iteration m + n deviates of the desired type were generated from distributions where $\theta_x = \theta_y$. The four tests were performed at each interation testing $H_o : \theta_x = \theta_y$ vs. $H_a : \theta_x > \theta_y$. The proportion of the iterations where the null hypothesis was rejected was recorded for each of the four tests. This proportion is the empirical level estimate. The

empirical levels were mulitplied by 1,000 and these values are reported in Table 3. Table 3 lists the empirical levels for each of the five distributions, for five sample size combinations, at each of the five variance ratios, τ, for each of the four tests. The standard error was calculated assuming a true nominal level of 0.05. We indicate an empirical level more than two standard deviations above 0.05 by entering the number into the table in boldface type.

Table 3. Empirical Levels Times 1,000 for $\alpha = 0.05$ for Each of the 4 Tests.

		m = 5, n = 5				m = 12, n = 5				m = 11, n = 10			
Distribution	τ	W	$\hat{U}^f$	B	t_s	W	$\hat{U}^f$	B	t_s	W	$\hat{U}^f$	B	t_s
Normal	0.1	49	49	47	51	22	35	49	48	**59**	**54**	49	50
	0.25	49	49	49	48	24	38	49	49	**55**	**54**	52	51
	1	51	51	**55**	50	40	52	**56**	49	51	53	52	50
	4	52	52	53	51	**67**	**61**	**56**	52	**55**	52	49	47
	10	52	52	48	51	**89**	**59**	52	53	**61**	**55**	48	45
Contaminated Normal	0.1	49	32	44	25	28	13	**59**	**73**	**59**	29	49	41
	0.25	48	29	47	21	29	13	**57**	**75**	**54**	25	49	38
	1	49	28	53	21	43	21	**58**	**73**	49	26	51	41
	4	45	27	46	22	**60**	27	49	**72**	**56**	29	52	40
	10	49	30	44	21	**76**	28	44	**72**	**61**	29	51	38
Uniform	0.1	50	50	47	**57**	19	32	47	46	**61**	52	47	50
	0.25	52	52	53	**55**	23	37	50	51	**55**	53	50	50
	1	47	47	49	44	42	**54**	**56**	**55**	48	50	49	49
	4	48	48	51	52	**80**	**66**	**59**	**59**	**62**	**56**	49	49
	10	49	49	48	**57**	**103**	**58**	53	**59**	**70**	**58**	50	50
Double Exponential	0.1	51	51	46	47	22	36	50	49	**58**	**54**	50	49
	0.25	49	49	44	47	25	39	50	46	49	50	48	48
	1	48	48	51	45	42	52	53	48	50	52	51	48
	4	50	50	51	47	**66**	**61**	**56**	47	**56**	53	49	47
	10	46	46	42	42	**87**	**59**	52	47	**66**	**58**	52	50
Gamma	0.1	32	32	30	32	10	19	27	38	33	30	28	37
	0.25	35	35	37	33	17	27	34	44	34	33	31	38
	1	48	48	51	46	42	53	**57**	**65**	51	53	51	50
	4	**87**	**87**	**88**	**85**	**123**	**110**	**100**	**109**	**129**	**122**	**115**	**84**
	10	**144**	**144**	**132**	**147**	**238**	**165**	**147**	**163**	**253**	**229**	**209**	**124**

Notes: Wilcoxon-Mann-Whitney (W), Fitted Test ($\hat{U}^f$), Brunner-Munzel (B), and Satterthwaite's t-test (t_s). Variance of X = 1, Variance of Y = τ. The right side of this table continues on the page below.

Table 3, continued.

Distribution	τ	m = 25, n = 20				m = 40, n = 40			
		W	$\hat{U}^f$	B	t_s	W	$\hat{U}^f$	B	t_s
Normal	0.1	51	47	48	48	**60**	44	48	49
	0.25	45	44	46	49	**54**	43	48	46
	1	47	46	48	50	52	46	52	53
	4	**59**	47	47	47	**58**	47	51	52
	10	**71**	50	49	48	**66**	49	52	52
Contaminated Normal	0.1	53	17	48	**64**	**62**	19	51	**54**
	0.25	50	16	49	**67**	**55**	17	50	**54**
	1	48	18	49	**67**	50	15	50	48
	4	**55**	19	48	**66**	52	16	49	52
	10	**65**	22	49	**66**	**63**	20	**54**	53
Uniform	0.1	**58**	48	49	49	**67**	44	48	47
	0.25	50	46	48	49	**61**	47	50	51
	1	50	49	51	51	51	46	52	50
	4	**63**	48	47	48	**61**	47	50	51
	10	**78**	52	50	52	**64**	43	46	46
Double Exponential	0.1	52	46	48	48	**64**	47	52	51
	0.25	48	46	49	49	51	43	47	48
	1	52	50	53	53	47	43	48	47
	4	**59**	48	48	47	**55**	46	50	**54**
	10	**70**	50	48	46	**61**	45	49	49
Gamma	0.1	22	19	20	40	20	13	15	41
	0.25	25	24	25	44	19	15	17	40
	1	47	46	48	50	52	48	53	**54**
	4	**186**	**158**	**158**	**76**	**245**	**215**	**227**	**72**
	10	**408**	**339**	**332**	**108**	**590**	**521**	**535**	**91**

Notes. Continued from previous page. Wilcoxon-Mann-Whitney (W), Fitted Test ($\hat{U}^f$), Brunner-Munzel (B), and Satterthwaite's t-test (t_s). Variance of X = 1, Variance of Y = τ.

There are a number of interesting conclusions that can be made from observing the values in Table 3. First, the t test using Satterthwaite's approximation for the degrees of freedom maintained its level when the data were generated from a normal distribution regardless of the sample size combination or the ratio of the variances. This was expected since the primary purpose of this test is to handle these situations. However, when the condition of normality was removed, the test became less predictable. In some cases, such as when the data were generated from the contaminated normal distribution and the sample sizes were similar, the test was very conservative. In other cases, such as when the data were uniformly distributed and when the sample sizes differed, the test became anti-conservative. The Wilcoxon-Mann-Whitney text generally does not maintain its level, even under the optimal condition of normality. It was conservative in situations where the larger sample size was taken from the population with the larger variance, and it was anti-conservative if the reverse was true. The fitted test generally maintained its level. In most of the situations

when it did not, the test was conservative. The Brunner-Munzel test generally maintained its level under all scenarios tested. All of the tests had trouble maintaining the 0.05 level when the data were simulated using the gamma distribution.

To estimate the tests' power, we ran 1,540 simulation iterations. This number of simulations assures that a 95% confidence interval for power will be approximately ± 0.025 when power is around 80%. For each iteration, m + n deviates of the desired type were generated under the condition $\theta_x - \theta_y = \delta$, where $\delta = \{1, 2, 3, 4\}$. Again, the proportion of the iterations where the null hypothesis was rejected was recorded for each of the four tests. This proportion is the test's estimated power. Since the Wilcoxon-Mann-Whitney test was anti-conservative in most scenarios, it was not surprising that the power of this test was greater than the power of the other tests. However, since this power is meaningless in the presence of an inflated nominal level, the Wilcoxon test will be removed from the rest of the discussion.

Figure 2 shows the power of the remaining tests under normality when the variances are not equal and the sample sizes are the same. Under these conditions, most statisticians would use the t test with Satterthwaite's approximation for the degrees of freedom. However, the fitted test and the Brunner-Munzel test demonstrate comparable power.

Figure 3 illustrates the power of the tests under normality with the added complication that the smaller sample size corresponds to the group with the larger variance. Once again, all three tests demonstrate similar power levels.

Figure 2. Plot of the Power for the Various Tests under Normality, Equal Sample Sizes, Ratio of the Variances $\tau = 0.1$, and $\alpha = 0.05$.

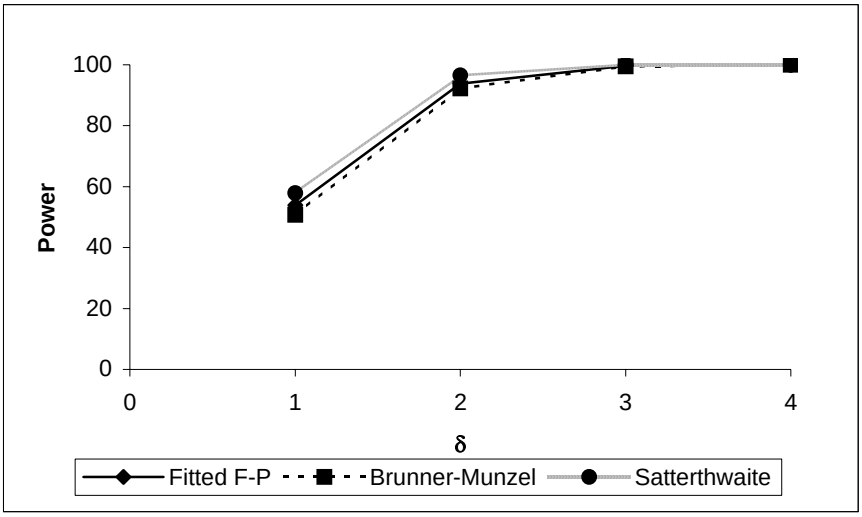
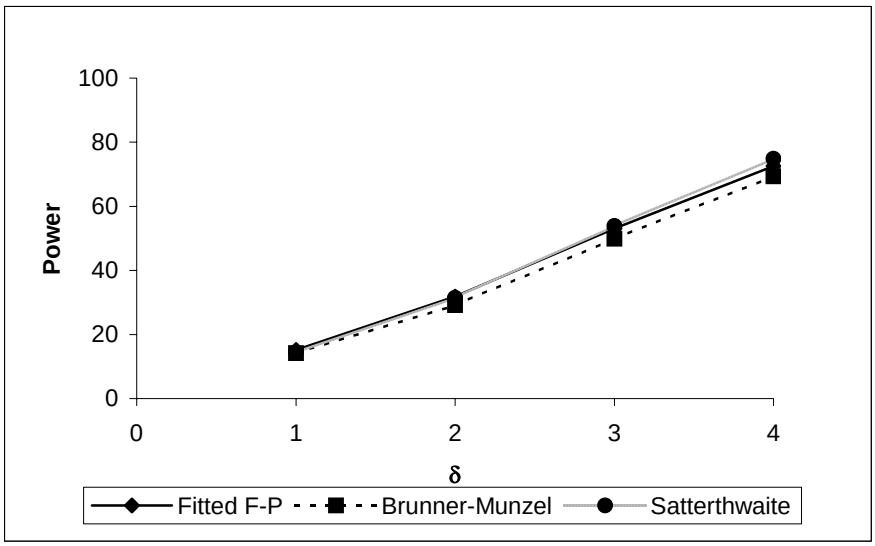


Figure 3. Plot of the Power for the Various Tests under Normality, Different Sample Sizes, Ratio of the Variances $\tau = 10$, and $\alpha = 0.05$.



When symmetry is removed from the distribution, such as in the case of the contaminated normal distribution, the fitted test and the Brunner-Munzel test demonstrate superiority over Satterthwaite's t test. This is illustrated in Figures 4 and 5. Figure 4 illustrates the power of the three tests when samples of the same size are generated from contaminated normal distributions with the same variance. Figure 5 illustrates the power of the three tests when samples of different sizes are generated from contaminated normal distributions with different variances. In both of these figures, the fitted test and the Brunner-Munzel test demonstrate similar power. However, the t test using Satterthwaite's approximation for the degrees of freedom has considerably less power than the other tests. This pattern is consistent over all of the results run using the contaminated normal distribution.

Figure 4. Plot of the Power for the Various Tests under the Contaminated Normal Distribution, Equal Sample Sizes, and Ratio of the Variances $\tau = 1$, and $\alpha = 0.05$.

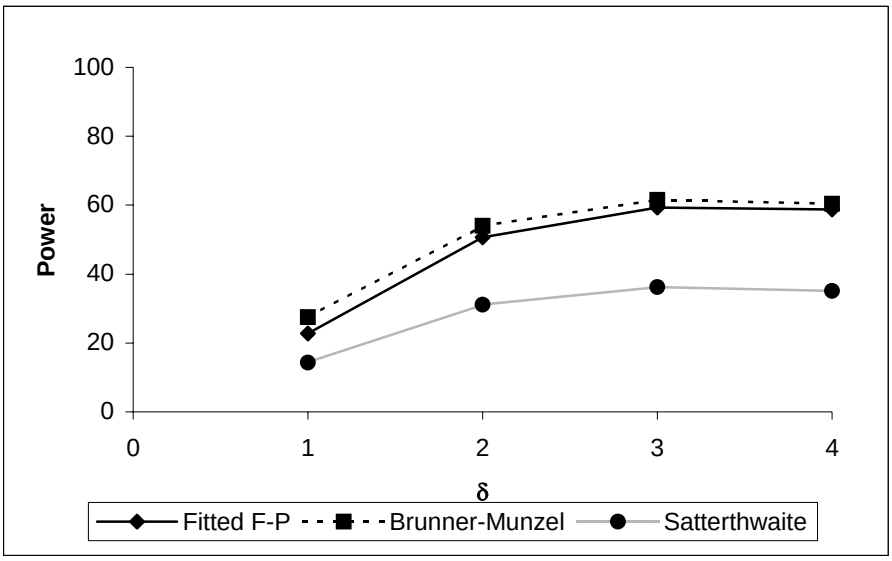
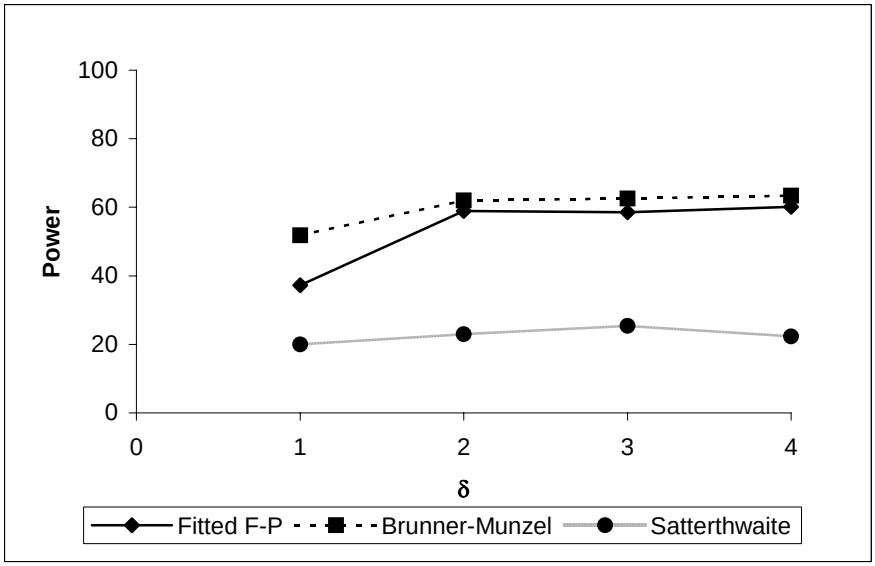


Figure 5. Plot of the Power for the Various Tests under the Contaminated Normal Distribution, Different Sample Sizes, and Ratio of the Variances $\tau = 0.1$, and $\alpha = 0.05$.



All three tests exhibited comparable power under the double exponential and uniform distributions. All three tests had increased power when the sample with fewer observations was obtained from the distribution with the smaller variance. However, all three tests exhibited decreased power when the sample with fewer observations was obtained from the distribution with the larger variance.

Conclusion

In this paper, we developed a method to test the difference between two population medians. Our fitted test was created by fitting a function to the large table of critical values presented by Fligner and Policello. Through a simulation study, we have determined that our test, and the Brunner-Munzel test, generally maintains the expected level of the test for a variety of underlying density functions.

The usual alternative to Student's t test, the Wilcoxon-Mann-Whitney test, has been shown to be anti-conservative in the simulation study under unequal variances by exhibiting empirical level estimates that are generally greater than the nominal level. Therefore, this test also exhibited artificially high power in the simulation results. Whereas, the fitted test and the Brunner-Munzel test have shown comparable power to the t test using Satterthwaite's approximation for the degrees of freedom under the ideal condition of normality. When symmetry is removed from the distribution function, such as in the contaminated normal distribution, the fitted test and the Brunner-Munzel test have shown improved power over the t test using Satterthwaite's approximation for the degrees of freedom.

All three tests exhibited comparable power in the simulation studies when the data were simulated from the double exponential or the uniform distributions. Statisticians should consider using an alternative to the Wilcoxon-Mann-Whitney test when unequal variances are possible.

References

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and small-sample approximation. *Biometrical Journal*, *42*, 17-25.

Brunner, E., & Neumann, N. (1982). Rank tests for correlated random variables. *Biometrical Journal*, *24*, 373-389.

Brunner, E., & Nuemann, N. (1986). Two-sample rank tests in general models. *Biometrical Journal*, *28*, 395-402.

Fligner, M. A., & Policello, G. E. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association, 76,* 162-168.

Lee, A. F. S., & Fineberg, N .S. (1991). A fitted test for the Behrens-Fisher problem. *Communications in Statistics – Theory and Methods*, *20*, 653-666.

Lee, A. F. S., & 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, 70*, 933-941.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*, 50-60.

Martinez, J., & Iglewicz, B. (1984). Some properties of the Tukey g and h family of distributions. *Communications in Statistics – Theory and Methods*, *13*, 353-369.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80-83.