A Simulation Study on Tests for the Behrens-Fisher Problem

Behrens-Fisher Problemi İçin Kullanılan Testler Üzerine Bir Simülasyon Çalışması

Evren ÖZKİP,[a]
Berna YAZICI,[a]
Ahmet SEZER[a]

[a]Department of Statistics,
Anadolu University Faculty of Science,
Eskişehir

Geliş Tarihi/*Received:* 17.06.2014
Kabul Tarihi/*Accepted:* 13.08.2014

Yazışma Adresi/*Correspondence:*
Berna YAZICI
Anadolu University Faculty of Science,
Department of Statistics, Eskişehir,
TÜRKİYE/TURKEY
bbaloglu@anadolu.edu.tr

ABSTRACT Objective: The problem of testing the equality of two means from normal populations with unknown variances is known as Behrens-Fisher (BF) problem. The difficulty with the BF problem is that exact solutions are not available satisfactorily because nuisance parameters are present. The aim is to compare the different methods for BF problem. **Material and Methods:** Classical t-test, Welch-Satterthwaite (WS) test, Cochran and Cox (CC) test, Singh-Saxena-Srivastava (SSS) test based on the jackknife procedure and generalized p-value (GP) test are considered. A Monte Carlo simulation study is conducted to evaluate type I error probabilities and powers of these methods. **Results:** Simulation results showed that when variances are unequal, the classical t-test is not an appropriate test because its type I error rate poorly deviates from the nominal level, α=0,05. The WS test has satisfactory type I error rate regardless of sample sizes and unequal variances. Its power appears to be more powerful than the other test when simple sizes are moderate or large. The GP test seems to be very conservative for small sample sizes. The power of the CC test is smallest among the five tests. However, its type I error rate has satisfactory as long as sample sizes are large. When sample sizes are unequal the type I error rate of the SSS test is too conservative or too liberal. **Conclusion:** The SSS test appears to be less powerful than the GP and WS tests as long as simple sizes are moderate and large. It is concluded that the SSS test is not as good as it has been reported by Singh et al (2002).

Key Words: Behrens-Fisher problem; Power; Type I error probability; Generalized p-value; Welch-Satterthwaite test

ÖZET Amaç: Normal dağılan iki anakütlenin ortalamalarının karşılaştırılmasında varyansların bilinmemesi Behrens-Fisher (BF) problemi olarak adlandırılır. BF problemin zorluğu nuisance parametrelerin varlığında tam olasılıklı çözümlerin elde edilememesindendir. Amacımız BF problemi için farklı metotları karşılaştırmaktır. **Gereç ve Yöntemler:** Klasik t testi, Welch-Satterthwaite (WS) testi, Cochran ve Cox (CC) testi, jackknife sürecine dayalı Singh-Saxena-Srivastava (SSS) testi ve genelleştirilmiş p-değeri (GP) testi ele alınmıştır. Bu metotların I. tip hata olasılıkları ve güçlerinin karşılaştırılması için bir Monte Carlo simülasyon çalışması yapılmıştır. **Bulgular:** Simülasyon çalışması gösterdi ki varyanslar eşit olmadığı zaman klasik t-test I. tip hata oranı nominal seviyeden (α=0.05) uzaklaştığı için uygun değildir. WS test, örneklem büyüklüğü ve varyansların farklı olması durumundan bağımsız tatmin edici tip I. tip hata oranına sahiptir. Örneklem büyüklükleri orta ve büyük olduğu durumlarda onun gücü diğer testlerden daha güçlü olduğu görüldü. Küçük örneklem için GP test nominal seviyeden daha küçük olduğu anlaşıldı. CC testin gücü diğer beş test arasında en küçüktür. Ama onun I. tip hata oranı büyük örneklemlerde tatmin edicidir. Örneklem büyüklükleri eşit olmadığı zaman SSS testin I. tip hata oranı nominal seviyeden çok küçük veya çok büyüktür. **Sonuç:** SSS testin gücü örneklem büyüklüğü orta veya büyük olduğu durumlarda GP ve WS testten daha küçüktür. SSS testinin Singh ve ark.nın (2002) ifade ettiği kadar iyi olmadığı sonucuna ulaşılmıştır.

Anahtar Kelimeler: Behrens-Fisher problem; Güç; I. tip hata olasılığı; Genelleştirilmiş p-değer; Welch-Satterthwaite testi

Turkiye Klinikleri J Biostat 2014;6(2):59-66

One frequently occurring problem encountered by researchers and applied statisticians is the testing of difference between two population means. The problem of comparing two distribution F_1 and F_2 is one of oldest problems in statistics. When two independent samples are available, the goal may be to compare the means of the distributions, i.e.

$$H_0: \mu_1 = \mu_2 \text{ vs } H_A: \mu_1 \neq \mu_2$$

where μ_i is the expectation of F_i, i= 1,2. In this case, many different procedures are available depending on the assumptions the analyst is ready to make about the data. The usual t -test is the test of choice when the variable of interest is normally distributed and its variances are the same for both distributions. But if the assumption of homogeneity of variances is violated, the usual t-test is no more robust for unequal sample sizes. The type I error probability is severely affected.

The problem of testing the equality of two means from normal populations with unknown variances is known as Behrens-Fisher (BF) problem. The Behrens-Fisher problem has been well known since the early 1930's. In the literature associated with the Behrens-Fisher problem, there have been quite a few solutions proposed. One reason for its popularity is that there is no exact solution satisfying the classical criteria for good tests. For example, Fisher, Welch, Aspin, Cochran and Cox, Qin and Jing have all suggested different solutions.[1-7] The idea is an extension of the solution for testing the Behrens-Fisher problem in the presence of nuisance parameters, which was proposed by Tsui and Weerahandi using the concept of the generalized p-value.[8] Tsui and Tang apply the distributional property of the generalized p-value for the BF problem to multiple testing.[9] Kim and Cohen propose a review of fundamental concepts and applications used to address the Behrens-Fisher problem under fiducial, Bayesian, and frequentist approaches.[10] Singh et al. proposed a new test using Jackknife methodology.[11] Dong considers the empirical likelihood approach for BF problem.[12] Chang and Pal revisit BF problem and apply a newly developed 'Computational Approach Test' (CAT).[13]

In this paper, we compare five methods for the difference between means of two normal populations when equal variances assumptions may be violated. We consider five methods: classical t-test, Welch-Satterthwaite (WS) test by Welch (1938 and 1947) and Satterhwaite, Cochran and Cox (CC) test by Cochran and Cox, Singh-Saxena-Srivastava (SSS) test based on the jackknife procedure by Singh et al. and generalized p-value (GP) test by Tsui and Weerahadi.[3,5,8,11,14,15]

The paper is organized as follows. In Section 2 gives Behrens Fisher problem. Section 3 introduces the five methods. Section 4 presents the simulation results to compare the type I error rates and powers of proposed methods. Section 5 gives discussion.

THE BEHRENS-FISHER PROBLEM

In practice, scientists interest to provide statistical inference on the mean difference of the two groups. For example they may want to detect a clinically meaningful mean difference or to establish therapeutic equivalence or bioequivalence between two drugs. If variances are equal then statistical inference is well solved. But when there are unequal variances, no satisfactory approach is available. Now assume that independent samples are available from two normal populations as:

$$X_{i1}, \ldots, X_{in_i} \; iid \; N(\mu_i, \sigma_i^2), i = 1{,}2$$

where all the four parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ are unknown. Based on the above two independent samples, problem is to test

$$H_0: \mu_1 = \mu_2 \text{ vs } H_A: \mu_1 \neq \mu_2 \qquad (1)$$

First, we reduce the above data by sufficiency, and focus only on $\bar{X}_{i.} = \sum_{j=1}^{n_i} X_{ij}/n_i$, $S_i^2 = \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.}\right)^2, i = 1{,}2$ where

$$\bar{X}_{i.} \sim N(\mu_i, \sigma_i^2/n_i) \text{ and } S_i^2 \sim \sigma_i^2 \chi^2_{(n_i-1)}, \; i = 1{,}2 \qquad (2)$$

and all the four statistics are mutually independent. Let $\bar{x}_1, \bar{x}_2, s_1^2, s_2^2$ denote the observed values $\bar{X}_1, \bar{X}_2, S_1^2, S_2^2$, respectively. The inferences are to be based on the set of complete sufficient statistics whose distributions are given by equation (1).

METHODS CONSIDERED

Statistical methods available in the literature can be roughly divided into either exact methods or approximation methods. In this section, we will propose four approximation methods and one exact method about the difference in the two means $\theta = \mu_1 - \mu_2$.

CLASSICAL T-TEST

For testing the hypothesis about the equality of the means under the assumption that $\sigma_1 = \sigma_2 = \sigma_1$, we use the statistic given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s}} \tag{3}$$

where s^2 is the pooled unbiased estimate of σ^2. This statistic has t-distribution with $n_1 + n_2 - 2$ degrees of freedom. This test is robust under violation of the assumption of normality. But if the assumption of homogeneity of variances is violated, t-test is no more robust for unequal sample sizes. The type I error probability is severely affect.[16]

WELCH SATTERTHWAITE (WS) TEST

According to Welch and Satterthwaite, the distribution of statistic t_{WS} can be approximated by t-distribution but this test statistic should be interpreted with reference to a modified number of degrees of freedom given by

$$t_{WS} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_1}\right)^2}{n_2 - 1}} \tag{4}$$

The WS test was originally designed to take care of unequal variances situation when both distributions are normal.

COCHRAN-COX (CC) TEST

The most prominent test after the WS test is the one proposed by Cochran and Cox (1950).[5] The CC test is defined by

$$t_{CC} = \frac{t_1 \frac{s_1^2}{n_1} + t_2 \frac{s_2^2}{n_2}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5}$$

where t_1 and t_2 are the tabulated critical values of the t-statistic for $n_1 - 1$ and $n_2 - 1$ degrees of freedom respectively.

SINGH-SAXENA-SRIVASTAVA (SSS) TEST

Singh et al. proposed a new test using Jackknife methodology.[11] This test can be used as an alternative to the BF problem. This new test is defined by

$$t_{SSS} = n t_{m,(\alpha/2)} \tag{6}$$

where t_m has t-distribution with m degrees of freedom and

$$n = \left((s_1^2/n_1) + (s_2^2/n_2)\right) / \left((s_1^2/n_2) + (s_2^2/n_1)\right),$$
$$m = \left((s_1^2/n_1) + (s_2^2/n_2)\right)^2 / \left((s_1^2/n_2)^2/(n_1 - 1) + (s_2^2/n_1)^2/(n_2 - 1)\right)$$

This test has some similarity to WS test. Note that the positions of and have been interchanged.

GENERALIZED P-VALUE (GP) TEST

The idea is an extension of the solution for testing the Behrens-Fisher problem in the presence of nuisance parameters, which was proposed by Tsui and Weerahandi using the concept of the generalized *p*-value.[8] Their generalized *p*-value method has been successfully used to provide small sample solutions for many hypothesis testing problems when nuisance parameters are present. The generalized pivotal quantity T for $\theta = \mu_1 - \mu_2$ can be defin-ed by

$$T(X_1, X_2; x_1, x_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \bar{x}_2 - \bar{x}_1 - \frac{\bar{X}_2 - \bar{X}_1 - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sqrt{\frac{\sigma_1^2 s_1^2}{n_1 S_1^2} + \frac{\sigma_2^2 s_2^2}{n_2 S_2^2}}$$
$$= \bar{x}_2 - \bar{x}_1 - Z \sqrt{\frac{s_1^2}{U_1} + \frac{s_2^2}{U_2}} \tag{7}$$

where

$$Z = \frac{\bar{X}_2 - \bar{X}_1 - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \quad , \quad U_i = \frac{n_i S_i^2}{\sigma_i^2} \sim \chi^2_{(n_i - 1)}, \qquad i = 1,2$$

and $s_1, s_2, \bar{x}_1, \bar{x}_2$ are the observed values of $S_1, S_2, \bar{X}_1, \bar{X}_2$, respectively. Generalized *p*-value based on generalized pivotal quantity for testing hypothesis in (1);

$$p = Pr(T > t_{obs} \ \backslash \theta = 0) \tag{8}$$

where t_{obs} is the observed value of T. This test reject the H_0 in (1) if $p < \alpha$, where α is the nominal level of testing.

SIMULATION STUDY

This section provides simulation studies for type I error probabilities and powers of the five methods proposed in Section 3. In this study, two configuration factors were taken into account to evaluate the performances of type I error probabilities and powers; sample size and variance.

To obtain the results of classical t-test, WS, CC and SSS tests, we use simulation consisting of 50,000 runs for each of the sample size and parameter configurations. The Monte Carlo method is used for estimating the type I error rates and powers of the GP test.

To obtain GP test, we use a two-step simulation. First we generated 2500 observed $(\bar{x}_1, \bar{x}_2; s_1^2, s_2^2)$, vectors and used 5000 runs for each observed vector to estimate the p-value in (8).

The estimates of type I error rates of five tests are present in Table 1. We have the following numerical results.

1. As long as the variances were homogeneous, the classical t-test seems to have a type I error rate quite close to α, the nominal level. But its type I error rate poorly deviates from the nominal level for unequal sample size and heteroscedasticity.

2. The WS test has satisfactory type I error rate, regardless of sample sizes and unequal variances. The CC test seems to be very conservative, when sample sizes are small. The WS and CC tests have similar result when the sample sizes are large.

3. For unequal sample sizes, the SSS test is far worse than type I error rates of the WS, CC and GP tests. But its type I error rate close to when sample sizes are equal and large.

4. The GP test seems to be very conservative when sample sizes are small.

The powers of five tests are present in Table 2. We once again observe from this table that the classical t-test appears to be more powerful than the others test when the variances were homogeneous and sample sizes are small. The power of CC test is smallest among the five tests. The WS and GP tests exhibit similar power values when simple sizes are large. In most case, WS test appears to be more powerful than the GP and SSS tests.

DISCUSSION

One frequently occurring problem encountered by researchers and applied statisticians is the testing of difference between two population means. The Behrens-Fisher problem is testing the equality of means from two independent normal populations without the assumption of equality of variances. In the literature associated with the Behrens-Fisher problem, there have been quite a few solutions proposed in the past several decades. The difficulty with the BF problem is that exact solutions are not available satisfactorily because nuisance parameters are present. In this paper, a Monte Carlo simulation was performed to evaluate the performance of the proposed five tests for the BF problem under different scenarios. Simulation studies were conducted to compare the type I error probabilities and powers of these methods. Simulation results show that when variances are unequal, the classical t-test is not an appropriate test because its type I error rate poorly deviates from the nominal level. The GP test seems to be very conservative for small sample sizes. However, its type I error rate close to $\alpha = 0.05$ and its power performs satisfactorily when sample sizes are equal and large. In most case, the WS test appears to be more powerful than the GP and SSS tests. Also it can be concluded that, regardless the sample sizes, its type I error rate, close to nominal level. The power of CC test is smallest among the five tests. However, its type I error rate has satisfactory as long as sample sizes are large. The WS and CC tests have similar size as well as power when the sample sizes are large. When sample sizes are unequal the type I rate error rate of SSS test is either too conservative or too liberal. The SSS test appears to be less powerful than the GP

TABLE 1: Type I error probabilities for the BF problem when $\mu_1 = \mu_2 = 0$ and $\alpha = 0.05$.

σ_1^2, σ_2^2	t-test	WS	CC	SSS	GP	σ_1^2, σ_2^2	t-test	WS	CC	SSS	GP
(n_1, n_2) = (5.5)						(n_1, n_2) = (5.10)					
(1,1)	0.050	0.045	0.023	0.072	0.026	(1,1)	0.050	0.052	0.032	0.093	0.024
(1,2)	0.051	0.044	0.025	0.069	0.026	(1,2)	0.035	0.049	0.030	0.100	0.022
(1,3)	0.053	0.046	0.027	0.066	0.024	(1,3)	0.029	0.048	0.030	0.103	0.026
(1,4)	0.055	0.047	0.029	0.065	0.026	(1,4)	0.026	0.048	0.031	0.103	0.030
(1,5)	0.057	0.049	0.031	0.063	0.028	(1,5)	0.024	0.049	0.031	0.100	0.030
(n_1, n_2) = (10.5)						(n_1, n_2) = (10.10)					
(1,1)	0.050	0.047	0.031	0.094	0.034	(1,1)	0.051	0.048	0.036	0.063	0.030
(1,2)	0.071	0.049	0.036	0.080	0.030	(1,2)	0.052	0.049	0.037	0.063	0.032
(1,3)	0.085	0.049	0.038	0.074	0.034	(1,3)	0.053	0.050	0.039	0.062	0.036
(1,4)	0.095	0.049	0.041	0.069	0.036	(1,4)	0.054	0.050	0.040	0.061	0.036
(1,5)	0.101	0.050	0.042	0.066	0.034	(1,5)	0.055	0.050	0.041	0.060	0.036
(n_1, n_2) = (10.25)						(n_1, n_2) = (25.10)					
(1,1)	0.050	0.049	0.042	0.080	0.040	(1,1)	0.050	0.052	0.042	0.078	0.044
(1,2)	0.030	0.050	0.041	0.108	0.044	(1,2)	0.077	0.051	0.044	0.054	0.046
(1,3)	0.022	0.049	0.041	0.126	0.044	(1,3)	0.094	0.051	0.045	0.043	0.040
(1,4)	0.018	0.050	0.041	0.136	0.044	(1,4)	0.105	0.051	0.045	0.036	0.044
(1,5)	0.015	0.051	0.041	0.144	0.048	(1,5)	0.113	0.051	0.046	0.032	0.046
(n_1, n_2) = (25.25)						(n_1, n_2) = (25.50)					
(1,1)	0.050	0.050	0.045	0.057	0.050	(1,1)	0.049	0.050	0.048	0.060	0.054
(1,2)	0.050	0.051	0.046	0.057	0.046	(1,2)	0.033	0.050	0.047	0.085	0.058
(1,3)	0.051	0.051	0.046	0.057	0.050	(1,3)	0.026	0.050	0.047	0.101	0.048
(1,4)	0.052	0.051	0.047	0.056	0.044	(1,4)	0.023	0.050	0.047	0.111	0.042
(1,5)	0.052	0.051	0.047	0.057	0.046	(1,5)	0.020	0.050	0.047	0.119	0.040
(n_1, n_2) = (50.25)						(n_1, n_2) = (50.50)					
(1,1)	0.050	0.051	0.047	0.059	0.048	(1,1)	0.050	0.051	0.048	0.053	0.038
(1,2)	0.071	0.050	0.047	0.038	0.054	(1,2)	0.049	0.050	0.048	0.054	0.034
(1,3)	0.084	0.051	0.048	0.029	0.054	(1,3)	0.050	0.050	0.049	0.054	0.030
(1,4)	0.092	0.051	0.048	0.025	0.052	(1,4)	0.049	0.050	0.049	0.054	0.032
(1,5)	0.097	0.051	0.049	0.022	0.052	(1,5)	0.050	0.050	0.049	0.056	0.034
(n_1, n_2) = (50.100)						(n_1, n_2) = (100.50)					
(1,1)	0.050	0.051	0.050	0.056	0.040	(1,1)	0.051	0.049	0.049	0.055	0.050
(1,2)	0.034	0.052	0.049	0.082	0.042	(1,2)	0.072	0.051	0.050	0.034	0.050
(1,3)	0.026	0.053	0.050	0.099	0.048	(1,3)	0.083	0.050	0.050	0.025	0.052
(1,4)	0.022	0.052	0.049	0.109	0.050	(1,4)	0.091	0.050	0.050	0.020	0.052
(1,5)	0.020	0.052	0.049	0.118	0.052	(1,5)	0.096	0.050	0.050	0.018	0.050
(n_1, n_2) = (100.100)						(n_1, n_2) = (100.200)					
(1,1)	0.050	0.050	0.050	0.052	0.048	(1,1)	0.051	0.050	0.051	0.054	0.052
(1,2)	0.051	0.049	0.050	0.052	0.044	(1,2)	0.033	0.049	0.051	0.081	0.052
(1,3)	0.051	0.049	0.050	0.052	0.044	(1,3)	0.026	0.049	0.051	0.099	0.048
(1,4)	0.051	0.049	0.050	0.053	0.046	(1,4)	0.022	0.049	0.050	0.109	0.046
(1,5)	0.051	0.049	0.050	0.053	0.048	(1,5)	0.020	0.049	0.050	0.118	0.046
(n_1, n_2) = (200.100)						(n_1, n_2) = (200.200)					
(1,1)	0.050	0.048	0.050	0.053	0.050	(1,1)	0.050	0.048	0.051	0.052	0.050
(1,2)	0.070	0.049	0.050	0.031	0.050	(1,2)	0.051	0.049	0.051	0.052	0.044
(1,3)	0.082	0.049	0.050	0.023	0.048	(1,3)	0.050	0.049	0.051	0.052	0.032
(1,4)	0.089	0.049	0.050	0.018	0.048	(1,4)	0.050	0.049	0.051	0.052	0.030
(1,5)	0.094	0.049	0.051	0.015	0.048	(1,5)	0.051	0.049	0.051	0.052	0.030

t-test: Classical t-test, WS: Welch-Satterthwaite test, CC: Cochran-Cox test, SSS: Singh-Saxena-Srivastava test, GP: Generalized p-value approach

TABLE 2: Powers for BF problem when nominal α = 0.05.

σ_1^2, σ_2^2	(μ_1, μ_2)	t-test	WS	CC	SSS	GP	t-test	WS	CC	SSS	GP
		(n_1, n_2) = (5.5)					(n_1, n_2) = (5.10)				
(1,1)	(0.3,0)	0.114	0.108	0.032	0.099	0.047	0.129	0.127	0.049	0.132	0.080
	(0.6,0)	0.219	0.208	0.075	0.176	0.113	0.272	0.259	0.119	0.239	0.167
	(0.9,0)	0.366	0.351	0.150	0.297	0.207	0.464	0.441	0.238	0.400	0.323
	(1.2,0)	0.533	0.517	0.260	0.448	0.349	0.663	0.631	0.401	0.578	0.503
	(1.5,0)	0.696	0.680	0.403	0.608	0.500	0.825	0.795	0.580	0.737	0.675
(1,0.5)	(0.3,0)	0.129	0.121	0.039	0.114	0.063	0.186	0.136	0.057	0.125	0.091
	(0.6,0)	0.264	0.248	0.098	0.219	0.136	0.375	0.284	0.140	0.238	0.193
	(0.9,0)	0.445	0.425	0.202	0.380	0.269	0.599	0.480	0.280	0.403	0.364
	(1.2,0)	0.637	0.614	0.351	0.561	0.449	0.791	0.679	0.460	0.584	0.562
	(1.5,0)	0.800	0.779	0.523	0.730	0.625	0.918	0.833	0.642	0.742	0.741
		(n_1, n_2) = (10.5)					(n_1, n_2) = (10.10)				
(1,1)	(0.3,0)	0.129	0.124	0.050	0.130	0.087	0.158	0.152	0.072	0.117	0.122
	(0.6,0)	0.271	0.259	0.121	0.241	0.181	0.362	0.357	0.206	0.281	0.314
	(0.9,0)	0.468	0.442	0.239	0.400	0.305	0.613	0.604	0.423	0.522	0.536
	(1.2,0)	0.665	0.635	0.401	0.580	0.515	0.824	0.820	0.666	0.754	0.776
	(1.5,0)	0.826	0.795	0.582	0.738	0.695	0.942	0.939	0.853	0.905	0.918
(1,0.5)	(0.3,0)	0.116	0.150	0.059	0.175	0.105	0.184	0.181	0.089	0.139	0.146
	(0.6,0)	0.285	0.344	0.168	0.355	0.224	0.441	0.432	0.265	0.356	0.360
	(0.9,0)	0.525	0.588	0.355	0.588	0.462	0.721	0.714	0.538	0.638	0.668
	(1.2,0)	0.749	0.798	0.583	0.790	0.692	0.907	0.904	0.793	0.859	0.870
	(1.5,0)	0.901	0.926	0.784	0.915	0.857	0.981	0.978	0.936	0.963	0.966
		(n_1, n_2) = (20.10)					(n_1, n_2) = (10.20)				
(1,1)	(0.3,0)	0.189	0.188	0.098	0.151	0.156	0.187	0.182	0.095	0.147	0.160
	(0.6,0)	0.447	0.439	0.284	0.362	0.418	0.444	0.437	0.281	0.360	0.388
	(0.9,0)	0.732	0.722	0.557	0.636	0.700	0.732	0.717	0.558	0.636	0.686
	(1.2,0)	0.915	0.904	0.808	0.850	0.884	0.915	0.905	0.808	0.851	0.894
	(1.5,0)	0.983	0.977	0.942	0.956	0.966	0.984	0.980	0.943	0.957	0.974
(1,0.5)	(0.3,0)	0.183	0.233	0.128	0.231	0.210	0.262	0.203	0.110	0.128	0.186
	(0.6,0)	0.501	0.578	0.406	0.557	0.556	0.575	0.490	0.329	0.345	0.440
	(0.9,0)	0.820	0.865	0.746	0.848	0.844	0.845	0.774	0.632	0.635	0.762
	(1.2,0)	0.966	0.977	0.940	0.970	0.966	0.968	0.940	0.866	0.858	0.932
	(1.5,0)	0.996	0.998	0.993	0.997	0.996	0.996	0.991	0.969	0.963	0.990
		(n_1, n_2) = (25.25)					(n_1, n_2) = (25.50)				
(1,1)	(0.3,0)	0.275	0.272	0.169	0.194	0.250	0.331	0.334	0.216	0.243	0.310
	(0.6,0)	0.672	0.672	0.529	0.568	0.634	0.784	0.782	0.660	0.685	0.780
	(0.9,0)	0.932	0.933	0.865	0.885	0.932	0.976	0.978	0.945	0.950	0.978
	(1.2,0)	0.993	0.994	0.983	0.987	0.996	0.999	1.00	0.998	0.998	1.00
	(1.5,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(1,0.5)	(0.3,0)	0.331	0.330	0.211	0.240	0.306	0.444	0.466	0.250	0.215	0.360
	(0.6,0)	0.778	0.778	0.653	0.688	0.776	0.884	0.922	0.735	0.685	0.848
	(0.9,0)	0.976	0.977	0.942	0.952	0.978	0.993	0.998	0.972	0.958	0.990
	(1.2,0)	0.999	1.00	0.998	0.998	1.00	1.00	1.00	0.999	0.999	1.00
	(1.5,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

continued →

TABLE 2: Powers for BF problem when nominal α = 0.05. *(continued)*

(1,1)	(0.3,0)	0.337	0.344	0.219	0.246	0.306	0.440	0.453	0.312	0.329	0.444
	(0.6,0)	0.784	0.790	0.659	0.686	0.786	0.910	0.913	0.838	0.849	0.904
	(0.9,0)	0.977	0.976	0.944	0.949	0.970	0.997	0.997	0.994	0.994	1.00
	(1.2,0)	1.00	1.00	0.998	0.998	1.00	1.00	1.00	1.00	1.00	1.00
	(1.5,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(1,0.5)	(0.3,0)	0.365	0.456	0.305	0.406	0.448	0.533	0.545	0.396	0.415	0.524
	(0.6,0)	0.871	0.910	0.832	0.888	0.916	0.963	0.961	0.924	0.931	0.966
	(0.9,0)	0.995	0.997	0.992	0.997	0.998	0.999	1.00	0.999	0.999	1.00
	(1.2,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(1.5,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		(n_1, n_2) = (50.100)					(n_1, n_2) = (100.50)				
(1,1)	(0.3,0)	0.533	0.543	0.400	0.415	0.560	0.531	0.531	0.399	0.416	0.522
	(0.6,0)	0.964	0.961	0.926	0.929	0.960	0.963	0.964	0.925	0.928	0.962
	(0.9,0)	1.00	1.00	0.999	0.999	1.00	1.00	1.00	1.00	0.999	1.00
	(1.2,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(1.5,0)	1.00	1.00	1.00	100	1.00	1.00	1.00	1.00	1.00	1.00
(1,0.5)	(0.3,0)	0.661	0.602	0.464	0.393	0.612	0.606	0.678	0.553	0.645	0.684
	(0.6,0)	0.990	0.982	0.960	0.940	0.986	0.991	0.994	0.986	0.993	0.990
	(0.9,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(1.2,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(1.5,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		(n_1, n_2) = (100.100)					(n_1, n_2) = (200.200)				
(1,1)	(0.3,0)	0.682	0.682	0.558	0.567	0.685	0.911	0.911	0.847	0.850	0.934
	(0.6,0)	0.995	0.995	0.987	0.988	0.992	1.00	1.00	1.00	1.00	1.00
	(0.9,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(1.2,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(1.5,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(1,0.5)	(0.3,0)	0.787	0.787	0.681	0.690	0.790	0.965	0.965	0.930	0.932	0.975
	(0.6,0)	0.999	0.999	0.999	0.999	0.997	1.00	1.00	1.00	1.00	1.00
	(0.9,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(1.2,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(1.5,0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

t-test: Classical t-test, WS: Welch-Satterthwaite test, CC: Cochran-Cox test, SSS: Singh-Saxena-Srivastava test, GP: Generalized p-value approach.

and WS tests for moderate and large samples. The SSS test is not as good as it has been reported by Singh et al.

Acknowledgement

This study is supported by the Research Foundation of Anadolu University. (Project Number: 1202F38).

REFERENCES

1. Fisher RA. The fiducial argument in statistical inference. Annals of Eugenics 1935;6(4):391-8.
2. Fisher RA. The asymptotic approach to Behrens' integral with further tables for the d test of significance. Annals of Eugenics 1941;11(1):141-72.
3. Welch BL. The generalisation of student's problems when several different population variances are involved. Biometrika 1947;34(1-2):28-35.
4. Aspin AA. An examination and further development of a formula arising in the problem of comparing two mean values. Biometrika 1948;35(Pts 1-2):88-96.
5. Cochran WG, Cox GM. Completely Randomized, Randomized Block, and Latin Square Designs. Chapter 4. Experiment Designs. 1st ed. New York: John Wiley and Sons; 1950. p.100-3.
6. Qin J. Likelihood and empirical likelihood ratio confidence intervals in two sample semi-parametric models. Madison: Technical Report Series University of Waterloo Stat; 1991. p.91-6.
7. Jing BY. Two-sample empirical likelihood method. Statistics and Probability Letters 1995;24(4):315-19.
8. Tsui KW, Weerahandi S. Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. J Amer Statist Assoc 1989;84(406):602-7.
9. Tsui KW, Tang SJ. Distributional property of the generalized p-value for the Behrens-Fisher problem with applications to multiple testing. Madison: University of Wisconsin; 2005. p.1-24.
10. Kim SH, Cohen AS. On the Behrens-Fisher problem: A review. Journal of Educational and Behavioral Statistics Winter 1998;23(4):356-77.
11. Singh P, Saxena KK, Srivastava OP. Power comparisons of solutions to the Behrens-Fisher problem. Am J Math Manag Sci 2002;22(3-4):233-50.
12. Dong LB. The Behrens-Fisher problem: An Empirical Likelihood Approach. Victoria, B.C: Econometrics Working Paper; 2004. p.1-29.
13. Chang CH, Pal N. A revisit to the Behrens–Fisher problem: Comparison of five test methods communications in statistics. Simulation and Computation 2008;37(7):1064-85.
14. Welch BL. The significance of the difference between two means when the population variances are unequal. Biometrika 1938;29(3-4): 350-62.
15. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics 1946;2(6):110-4.
16. Hsu PL. Contribution to the theory of "student's" t-test as applied to the problem of two samples. Statistical Research Memoirs 1938; 2(1):1-24.