American Journal of Mathematics and Statistics 2023, 13(1): 60-68
DOI: 10.5923/j.ajms.20231301.03

Parametric Solutions to the Behrens-Fisher Problem

Ibrahim H. Ibrahim, Ghada Taha*, Mahmoud Sadek

Department of Mathematics, Insurance and Applied Statistic, Helwan University, Cairo, Egypt

Abstract The Behrens-Fisher (B-F) problem arises from testing the equality between two population means from independent normal populations when variances are unknown, and the variances cannot assume to be equal. Many literatures have been introduced to solve this problem and several solutions have been proposed for it. In this article, two tests are proposed to deal with the B-F problem. These two proposed tests depended on the test statistic that was introduced by Behrens (1929) with some modifications that are based on the method that was provided by Chen et al. (2022) which depended on Fisher's fiducial argument to estimate the variances of the sample means. Also, the formula for degree of freedom and constant were derived for each suggested solution. The comparison among the proposed tests and some existing tests such as Welch test and Fenstad test have been studied extensively by Monte Carlo simulation. The size and the power of these tests are evaluated by using several simulation scenarios to assess the suggested tests. The comparison study proved that the sample sizes and variances of populations should be taken into consideration to decide which tests should be used when dealing with this problem. This study shows that the power of proposed tests are better than or close to the power of Welch test especially, when the sample sizes are large regardless of this data is balanced or unbalanced.

Keywords Behrens-Fisher problem, Welch test, Fenstad test, Fisher's fiducial argument, Power of the test, Size of the test, Balanced data, Unbalanced data

1. Introduction

The Behrens-Fisher (B-F) problem occurs when testing the equality between two population means from independent normal populations when variances are unknown, and the variances cannot assume to be equal [13] [9].

Several solutions introduced and developed to solve this problem. Behrens (1929) proposed the earliest solution to this problem, Fisher (1939) endorsed this solution. Therefore, this problem is known Behrens-Fisher (B-F) Problem. But, this solution was not acceptable to many statisticians because the size or the estimated type I error of this test is often less than the nominal level [15] [8] [3] [1].

Ever since, several solutions proposed for this problem and there was no exact solution to satisfy for all sample sizes [3] [11]. The popular approximation solution proposed by Welch (1938). Also, various approximation solutions were proposed such as: Cochran approximations' which depends on the Behrens- Fisher test statistic with different degrees of freedom [5]. On the other hand, another solution proposed by Fenstad (1983). However, Fenstad did not derive the formula of degrees of freedom for this test statistic that was an approximation to the t-distribution. Best and Rayner (1987) derived the degree of freedom formula for Fenstad test, and Paul (1992) showed that there exists an error in the degree of freedom formula. Best and Rayner (1987) proposed other solutions for this problem as the score test and Wald test. Modified Mover test is one of the latest solutions that proposed By Chen et al. (2022).

Based on empirical properties of two proposed tests that deal with B-F problem, this paper aims to compare these two proposed tests with some existing tests such as Welch test and Fenstad test according to two comprehensive Monte Carlo simulation studies with different scenarios on the size and the power of the test. This simulation study was based on three factors (i) sample sizes (balanced or unbalanced), (ii) variances of populations, and (iii) the gap between population variances. This paper proceeds as follow. Section literature Survey for the Behrens- Fisher problem. Then, the proposed solutions were presented in section 3. Section 4 demonstrated the simulation study. Finally, the conclusion of the study is shown in section 5.

2. Literature Survey

Several solutions were proposed for B-F problem. These solutions can be classified into exact and approximated solutions. In this paper, we focused on three approximation solutions (i) the Welch test, (ii) the Fenstad test, and (iii) the Wald test [11] [14].

For testing the equality between two population means when variances are unknown or unequal based on two independent samples $x_1,\ldots, x_n$ and $y_1,\ldots, y_m$ from $N(\mu_1,\sigma_1^2)$

* Corresponding author:
ghadataha@commerce.helwan.edu.eg (Ghada Taha)
Received: Sep. 29, 2023; Accepted: Oct. 22, 2023; Published: Oct. 28, 2023
Published online at http://journal.sapub.org/ajms

and $N(\mu_2,\sigma_2^2)$ respectively; where $-\infty<\mu_k<\infty$ and $0<\sigma_k^2<\infty$ for k =1,2. The null and the alternative hypotheses are: $H_0: \mu_1=\mu_2$ (or $\mu_1-\mu_2=0$) vs $H_1: \mu_1>\mu_2$(or $\mu_1-\mu_2>0$) . First, we define some statistics as

$$\bar{x}=\sum_{i=1}^{n} x_i/n,\ \bar{y}=\sum_{j=1}^{m} y_j/m \tag{1}$$

$$S_1^2=\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1),\ S_2^2=\sum_{j=1}^{m}(y_j-\bar{y})^2/(m-1) \tag{2}$$

Where $\bar{x},\bar{y}$ are the sample means and S_1^2, S_2^2 are the sample variances for the first and second sample respectively. So that:

$$\bar{x}\sim N(\mu_1,\frac{\sigma_1^2}{n}),\ \text{and}\ \bar{y}\sim N(\mu_2,\frac{\sigma_2^2}{m}) \tag{3}$$

$$\bar{x}-\bar{y}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}) \tag{4}$$

Then

$$\frac{(n-1)s_1^2}{\sigma_1^2}\sim X^2(n-1),\ \frac{(m-1)s_2^2}{\sigma_2^2}\sim X^2(m-1) \tag{5}$$

Where, $X^2(k)$ is chi- square probability distribution with k degrees of freedom.

Therefore,

$$E\left(\frac{(n-1)s_1^2}{\sigma_1^2}\right)=n-1,\ \text{and}\ E\left(\frac{(m-1)s_2^2}{\sigma_2^2}\right)=m-1 \tag{6}$$

And thus,

$$E(s_1^2)=\sigma_1^2,\ E(s_2^2)=\sigma_2^2 \tag{7}$$

Therefore, S_1^2, S_2^2 are unbiased estimators for σ_1^2, σ_2^2 respectively.

Welch test (T1): This test was proposed by Welch (1938), this is well known as a standard solution to testing the equality between two means from normal population with unequal variances, [3] [4]. According to this test, the test statistic was approximated by t-distribution with degrees of freedom ($f_{(1)}$). We can calculate the Welch statistic T1 and $f_{(1)}$ as the following [14] [3]:

$$T1=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}} \tag{8}$$

$$f_{(1)}=\frac{\left(\frac{s_1^2}{n}+\frac{s_2^2}{m}\right)^2}{\left(\frac{\left(\frac{s_1^2}{n}\right)^2}{(n-1)}\right)+\left(\frac{\left(\frac{s_1^2}{m}\right)^2}{(m-1)}\right)} \tag{9}$$

Fenstad test (T2): Fenstad (1983) suggested a test statistic to deal with B-F problem as the following [3] [13]:

$$T2=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{(n-1)s_1^2}{n^2-3n}+\frac{(m-1)s_2^2}{m^2-3m}}} \tag{10}$$

Where, T2 was approximated by t-distribution with degrees of freedom ($f_{(2)}$) and constant ($C_{(2)}$) as the following:

$$T2\sim C_{(2)}t_{f_{(2)}} \tag{11}$$

$$f_{(2)}=\frac{\left(\frac{(n-1)s_1^2}{n(n-3)}+\frac{(m-1)s_2^2}{m(m-3)}\right)^2}{\frac{(n-1)s_1^4}{n^2(n-3)^2}+\frac{(m-1)s_2^4}{m^2(m-3)^2}} \tag{12}$$

$$C_{(2)}=\frac{\frac{s_1^2}{n}+\frac{s_2^2}{m}}{\frac{(n-1)s_1^2}{n(n-3)}+\frac{(m-1)s_2^2}{m(m-3)}} \tag{13}$$

$f_{(2)}$ Was proposed by Paul (1992) and $C_{(2)}$ was introduced by [3].

Wald test (W): This test was proposed by Best and Rayner (1987). The Wald test statistic is:

$$W=\frac{(\bar{x}-\bar{y})^2}{\left(\frac{(n-1)s_1^2}{n^2}+\frac{(m-1)s_2^2}{m^2}\right)} \tag{14}$$

Best and Rayner only suggested the formula of the test statistic without proposing the approximated distribution for it.

3. Proposed Solutions to the B-F Problem

In this paper, we suggest new two solutions to the B-F problem to solve the B-F problem. These solutions depended on the test statistic that introduced by Behrens (1929), supported by Fisher (1939) and used by Welch (1938) [10]. Also, we use the method that was provided by Chen et al. (2022) that based on Fisher's fiducial argument to estimate the variances of the sample means and substituting with them in the test statistic. Then, we derive the formula for degree of freedom and constant for each suggested solution. Let T be the test statistic:

$$T=\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}} \tag{15}$$

$$T=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{v(\bar{x})+v(\bar{y})}} \tag{16}$$

When H_0 is true (or $\mu_1-\mu_2=0$). The test statistic can be written as the following:

$$T=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}}=\frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\widehat{v(\bar{x})}+\widehat{v(\bar{y})}}} \tag{17}$$

Welch (1938) approximated this test statistic to the student t- distribution. It can be written as: $T\sim c\,t_f$, where f is the degrees of freedom, and c is a constant (c=1).

To get the values of the test statistic, we need to get:

$$v(\bar{x})+v(\bar{y})=\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m} \tag{18}$$

In the B-F problem, σ_k^2 is often unknown, we can replace it by using the variance estimate $\hat{\sigma}_k^2$. Therefore, we can rewritten equation (19) as:

$$\widehat{v(\bar{x})}+\widehat{v(\bar{y})}=\frac{\hat{\sigma}_1^2}{n}+\frac{\hat{\sigma}_2^2}{m} \quad (19)$$

Then σ_j^2 can be estimated by using the following relationships [4]:

$$\frac{(n-1)s_1^2}{\sigma_1^2}\sim X^2(n-1),\ \frac{(m-1)s_2^2}{\sigma_2^2}\sim X^2(m-1) \quad (20)$$

Let:

$$\frac{(n-1)s_1^2}{\sigma_1^2}=U_k,\quad k=1,2 \quad (21)$$

Where, U_k is a random variable that follows the chi-square X^2 distribution.

Therefore:

$$\sigma_1^2=\frac{s_1^2(n-1)}{U_1},\ \sigma_2^2=\frac{s_2^2(m-1)}{U_2} \quad (22)$$

We can assume some values for U_k to get the corresponding values of $\hat{\sigma}_k^2$. Different values for U_k will lead to different values of $\hat{\sigma}_k^2$.

Chen et al. (2022) replaced U_k with (n-3), which is the maximum value of probability density function when introduced the Modified Mover statistic. If we replaced U_k with (n-1), get the variance estimate and substitute with it in equation (8), we will get to the Welch statistic. Where, (n-1) is the mean of $\mathcal{X}^2$ distribution. Thus, in this paper we proposed two different cases, the values (n, m) will replace the variables (U_1, U_2) respectively as will be shown in the Case-I. While, in the Case-II we consider (n-2, m-2) to replace the variables (U_1, U_2) respectively.

For each suggested test statistic, we need to get the degrees of freedom (f) and the constant (C) to approximate the test statistic to t- distribution as we shown in Welch approximation. By investigating the previous solutions for the B-F problem which introduced in statistical literatures such as: Welch test and Fenstad test, we can derive the formulas for (f), (C) as:

$$f=\frac{(\grave{v}(\bar{x})+\grave{v}(\bar{y}))^2}{\left(\frac{(\grave{v}(\bar{x}))^2}{n-1}+\frac{(\grave{v}(\bar{y}))^2}{m-1}\right)} \quad (23)$$

$$C=\frac{v(\bar{x}_1)+v(\bar{y})}{\grave{v}(\bar{x})+\grave{v}(\bar{y})} \quad (24)$$

$$\widehat{v(\bar{x})}+\widehat{v(\bar{y})}=\frac{s_1^2}{n}+\frac{s_2^2}{m} \quad (25)$$

Where $\grave{v}(\bar{x}),\grave{v}(\bar{y})$ are variances values of the sample mean for the first and second sample respectively.

$v(\bar{x}_1),v(\bar{x}_2)$ are variances value of the sample mean which was used by Behrens and Fisher before for the first and second sample respectively.

3.1. Case-I: (U_1= n, U_2= m)

In this case, we need to get the variance estimate by replacing the variables (U_1, U_2) with (n, m) respectively. Then equation (22) can rewritten as the following:

$$\hat{\sigma}_1^2=\frac{(n-1)s_1^2}{U_1},\ \hat{\sigma}_2^2=\frac{(m-1)s_2^2}{U_2} \quad (26)$$

Therefore, the variance estimator is given as:

$$\hat{\sigma}_1^2=\frac{(n-1)s_1^2}{n},\ \hat{\sigma}_2^2=\frac{(m-1)s_2^2}{m} \quad (27)$$

$$\widehat{v(\bar{x})}=\frac{(n-1)s_1^2}{n^2},\ \widehat{v(\bar{y})}=\frac{(m-1)s_2^2}{m^2} \quad (28)$$

We can substitute with $\widehat{v(\bar{x})}$, $\widehat{v(\bar{y})}$ in equation (17) to get the new test statistic as follow:

$$T3=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{(n-1)s_1^2}{n^2}+\frac{(m-1)s_2^2}{m^2}}}=\sqrt{W} \quad (29)$$

Where $\sqrt{W}$ is the square root of Wald statistic, thus $T3\sim C_{(3)}t_{f_{(3)}}$.

By applying equations (23) and (24), respectively. We get $f_{(3)}, C_{(3)}$ as the following:

$$f_{(3)}=\frac{\left(\frac{(n-1)s_1^2}{n}+\frac{(m-1)s_2^2}{m}\right)^2}{\left(\frac{\left(\frac{(n-1)s_1^2}{n^2}\right)^2}{(n-1)}\right)+\left(\frac{\left(\frac{(m-1)s_2^2}{m^2}\right)^2}{(m-1)}\right)} \quad (30)$$

Then,

$$f_{(3)}=\frac{\left(\frac{(n-1)s_1^2}{n}+\frac{(m-1)s_2^2}{m}\right)^2}{\frac{(n-1)s_1^4}{n^4}+\frac{(m-1)s_2^4}{m^4}} \quad (31)$$

$$C_{(3)}=\frac{\frac{s_1^2}{n}+\frac{s_2^2}{m}}{\frac{(n-1)s_1^2}{n^2}+\frac{(m-1)s_2^2}{m^2}} \quad (32)$$

3.2. Case-II: (U_1=n-2, U_2=m-2)

In this case, we need to get the variance estimate by replacing the variables (U_1, U_2) with (n-2, m-2) respectively. Then, we can reformulate equation (12) as the following:

$$\hat{\sigma}_1^2=\frac{(n-1)s_1^2}{n-2},\ \hat{\sigma}_2^2=\frac{(m-1)s_2^2}{m-2} \quad (33)$$

$$\widehat{v(\bar{x})}=\frac{(n-1)s_1^2}{n(n-2)},\ \widehat{v(\bar{y})}=\frac{(m-1)s_2^2}{m(m-2)} \quad (34)$$

Then, we can substitute with $\widehat{v(\bar{x})}$, $\widehat{v(\bar{y})}$ in equation (17) to get the test statistic T4 as the following:

$$T4=\frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{(n-1)s_1^2}{n(n-2)}+\frac{(m-1)s_2^2}{m(m-2)}}} \quad (35)$$

This test statistic is approximated by t-distribution with degrees of freedom ($f_{(4)}$) and constant value ($C_{(4)}$) as:

$$T4\sim C_{(4)}t_{f_{(4)}} \quad (36)$$

By using the formulas in equation (23) & (24) we can get $f_{(4)}, C_{(4)}$ as the following:

$$f(4)=\frac{\left(\frac{(n-1)s_1^2}{n(n-2)}+\frac{(m-1)s_2^2}{m(m-2)}\right)^2}{\left(\frac{\left(\frac{(n-1)s_1^2}{n(n-2)}\right)^2}{(n-1)}\right)+\left(\frac{\left(\frac{(m-1)s_2^2}{m(m-2)}\right)^2}{(m-1)}\right)} \quad (37)$$

$$C(4)=\frac{\frac{s_1^2}{n}+\frac{s_2^2}{m}}{\frac{(n-1)s_1^2}{n(n-2)}+\frac{(m-1)s_2^2}{m(m-2)}} \quad (38)$$

4. Simulation Study

The Monte Carlo simulation study was conducted using R package as shown in the following steps:

1- Generating data for samples from normal populations at different combinations of the four factors that we referred to them in the previous section.
2- Calculate the estimated sample means ($\bar{x}$, $\bar{y}$) for each case.
3- Calculate the estimated variances ($\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$) for each case.
4- Calculate the test statistics for the four tests.
5- Calculate the size of the four tests.
6- Calculate the power of the four tests.

A comparative study was conducted to evaluate the performance of four tests:

(1) Welch test (T1),
(2) Fenstad test (T2), and the proposed tests:
(3) The first proposed test (T3), and
(4) The second proposed test (T4).

Table 1. The Probability of Type-I Error for The Four Tests Under Different Variances, $\mu_k = 2$ and n = m = 20

Var(1)	Var(2)	T1	T2	T3	T4
1	0.5	0.0499	0.0631	0.045	0.0564
5	3	0.0466	0.0608	0.0412	0.0536
8	2	0.0472	0.0601	0.041	0.0534
12	6	0.0508	0.0642	0.0468	0.0574
16	8	0.0505	0.0632	0.0455	0.0562
20	15	0.0498	0.0610	0.0449	0.0548
25	20	0.0463	0.0600	0.0421	0.0519
36	24	0.0455	0.0583	0.0417	0.0522
40	45	0.0526	0.0644	0.0475	0.0572
50	60	0.0526	0.0659	0.0469	0.0599
60	75	0.049	0.0632	0.0434	0.0562
70	100	0.0483	0.0617	0.0435	0.0549
60	80	0.0491	0.0638	0.0436	0.0561
80	100	0.0454	0.0571	0.0412	0.0503
85	120	0.0521	0.0634	0.0473	0.0572
105	150	0.0525	0.0638	0.0477	0.0582

These simulation studies are based on three factors: (i) sample sizes (balanced or unbalanced), (ii) values of the variances of the populations, and (iii) the gap between population variances. In several scenarios, the simulation studies were conducted to compare the size (The probability of type-I error) and the power of each test under different factors combinations. These simulation studies were applied with samples generated from normal populations with different means and different variances in two scenarios as the following:

Case 1: Balanced data (the sample sizes are equal).

Case 2: Unbalanced data (the sample sizes are different).

In Tables (1, 2 and 3), the estimated type-I error probabilities for the four tests are shown when the sample sizes are (n, m = 20, 50 and 100).

Table 2. The Probability of Type-I Error for The Four Tests Under Different Variances, $\mu_k = 2$ and n =m = 50

Var(1)	Var(2)	T1	T2	T3	T4
1	0.5	0.0514	0.0563	0.0487	0.0534
5	3	0.0511	0.0552	0.049	0.0532
8	2	0.0497	0.055	0.047	0.052
12	6	0.048	0.0522	0.0455	0.0505
16	8	0.0462	0.0501	0.0442	0.0481
20	15	0.0466	0.0534	0.0449	0.05
25	20	0.0474	0.0534	0.0452	0.0502
36	24	0.0494	0.054	0.0471	0.0524
40	45	0.0476	0.0524	0.0451	0.0502
50	60	0.0511	0.0556	0.0488	0.053
60	75	0.0462	0.0515	0.0437	0.049
70	100	0.0498	0.0541	0.0471	0.0521
60	80	0.0529	0.0572	0.0505	0.0556
80	100	0.0487	0.0522	0.0465	0.0505
85	120	0.0537	0.0584	0.052	0.056
105	150	0.0516	0.0559	0.0482	0.054

Table 3. The Probability of Type-I Error for The Four Tests Under Different Variances, $\mu_k = 2$ and n= m = 100

Var(1)	Var(2)	T1	T2	T3	T4
1	0.5	0.0511	0.0533	0.0495	0.0523
5	3	0.0500	0.0527	0.0484	0.0512
8	2	0.0551	0.0571	0.0538	0.0561
12	6	0.0513	0.0535	0.0503	0.0524
16	8	0.0493	0.0524	0.0482	0.0506
20	15	0.0494	0.0519	0.0481	0.0508
25	20	0.0521	0.0551	0.051	0.0535
36	24	0.0512	0.0539	0.0501	0.0527
40	45	0.0539	0.0562	0.0519	0.0554
50	60	0.0509	0.0533	0.0489	0.0517
60	75	0.0498	0.0529	0.0485	0.0512
70	100	0.0520	0.054	0.0504	0.053
60	80	0.0485	0.0503	0.0475	0.0496
80	100	0.0506	0.0539	0.0493	0.0524
85	120	0.0510	0.0539	0.0488	0.0523
105	150	0.0520	0.054	0.0504	0.0530

These simulation studies determined based on 10000 generated samples at a nominal level $\alpha = 0.05$, the samples generated from normal distribution at $\mu = 2$ under different variances. Var(1) and Var(2) are the variances of the first and the second population respectively. We used the R package for the computations in this paper.

Figure. 1 represents the estimated type-I error probabilities (transformed by log10) for the four tests. This figure corresponding to values of Tables (1, 2 and 3). Two vertical lines in this figure represent the solid and broken lines equivalent to 0.05 and 0.06, respectively as shown in Figures. 1 and 3. Different symbols in these figures represent the different sample sizes (n, m= 20, 50 and 100) as shown in Figures. 1, 2, 3 and 4.

Figure. 1 demonstrates that the estimated type-I error for test T1 (Welch test) is closer to a nominal probability at 0.05 in all combinations (acceptable size). But test T2 has overestimated probably of type-I error when sample sizes are small regardless of the values of population variances. Therefore, test T2 cannot be recommended for testing the differences between the two means in this case.

The estimated type-I error for test T3 is so far from a nominal level at 0.05 when the sample sizes are small and variances are small. But it becomes closer to 0.05 when increasing sample size and the value of variances.

type-I error for test T4 is between the two lines, that is mean that the size of this test is acceptable.

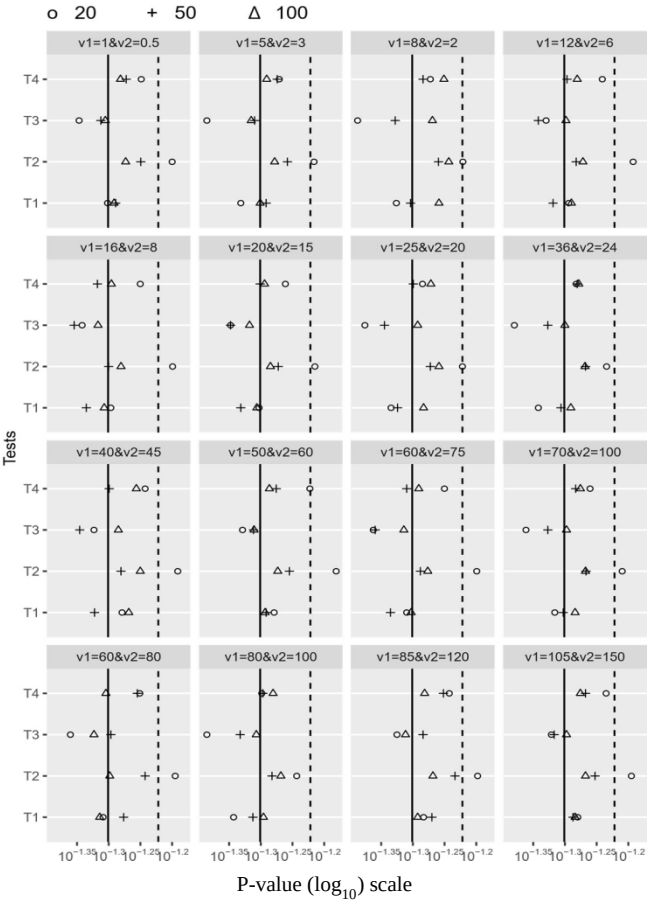

Figure 1. The estimated probabilities of type-I error for the four tests

Tables (4, 5 and 6) represent the power of the four tests when the sample sizes are equal (balanced data) and ($\mu_1 = 2$, $\mu_2 = 8$) under different variances.

Table 4. The Power of The Test for The Four Tests Under Different Variances, $\mu_1 = 2$, $\mu_2 = 8$ and n=m=20

Var(1)	Var(2)	T1	T2	T3	T4
1	0.5	99.92%	94.84%	99.93%	99.99%
5	3	99.87%	94.04%	99.87%	99.95%
8	2	99.83%	93.73%	99.83%	99.89%
12	6	99.77%	93.30%	99.79%	99.83%
16	8	99.63%	92.75%	99.63%	99.72%
20	15	99.01%	91.79%	98.94%	99.15%
25	20	97.16%	90.17%	96.84%	97.54%
36	24	91.85%	85.68%	91.01%	92.57%
40	45	80.54%	76.11%	79.13%	81.98%
50	60	70.27%	66.81%	68.39%	71.97%
60	75	61.13%	58.81%	59.13%	63.15%
70	100	50.89%	49.57%	49.00%	53.00%
60	80	59.21%	56.96%	57.12%	61.48%
80	100	49.31%	47.44%	47.46%	51.48%
85	120	43.90%	42.83%	41.85%	46.13%
105	150	36.59%	35.81%	34.68%	38.42%

Table 5. The Power of The Test for The Four Tests Under Different Variances, $\mu_1 = 2$, $\mu_2 = 8$ and n= m= 50

Var(1)	Var(2)	T1	T2	T3	T4
1	0.5	99.93%	94.64%	99.92%	99.97%
5	3	99.86%	94.47%	99.87%	99.90%
8	2	99.85%	93.75%	99.84%	99.86%
12	6	99.76%	93.09%	99.79%	99.82%
16	8	99.73%	92.76%	99.73%	99.77%
20	15	99.65%	92.50%	99.69%	99.72%
25	20	99.63%	91.85%	99.62%	99.67%
36	24	99.55%	91.29%	99.55%	99.57%
40	45	99.12%	90.19%	99.04%	99.16%
50	60	97.64%	88.78%	97.53%	97.82%
60	75	94.30%	85.59%	94.14%	94.49%
70	100	88.90%	80.29%	88.53%	89.36%
60	80	93.64%	83.88%	93.40%	94.00%
80	100	86.97%	77.88%	86.62%	87.44%
85	120	82.47%	73.86%	81.89%	83.05%
105	150	74.17%	66.47%	73.61%	74.88%

Figure. 2 represents the estimated power of the test for the four tests. This figure corresponding to values of Tables (4, 5 and 6).

The power of the test for test T4 is better than the power for test T1 when the values of variances are small regardless of the sample sizes. Also, the power for the test T3 is high but lower than the power of T1 slightly. In general, the power for tests T1, T3 and T4 are decreasing with increasing the values of variances and gap of these variances.

Table 6. The Power of The Test for The Four Tests Under Different Variances, $\mu_1 = 2$, $\mu_2 = 8$ and n=m= 100

Var(1)	Var(2)	T1	T2	T3	**T4**
1	0.5	97.90%	94.53%	96.92%	99.98%
5	3	97.90%	94.03%	96.87%	99.94%
8	2	97.83%	93.62%	96.81%	99.86%
12	6	97.78%	93.50%	96.77%	99.81%
16	8	97.75%	92.75%	96.73%	99.77%
20	15	97.66%	92.28%	96.69%	99.74%
25	20	97.62%	91.60%	96.65%	99.66%
36	24	97.59%	91.40%	96.61%	99.61%
40	45	97.55%	90.81%	96.55%	99.59%
50	60	97.49%	90.38%	96.47%	99.52%
60	75	97.38%	89.88%	96.38%	99.38%
70	100	97.00%	89.02%	95.97%	99.00%
60	80	97.17%	88.61%	96.17%	99.22%
80	100	96.65%	87.67%	95.59%	98.67%
85	120	95.90%	86.43%	94.90%	97.95%
105	150	93.36%	83.69%	92.32%	95.37%

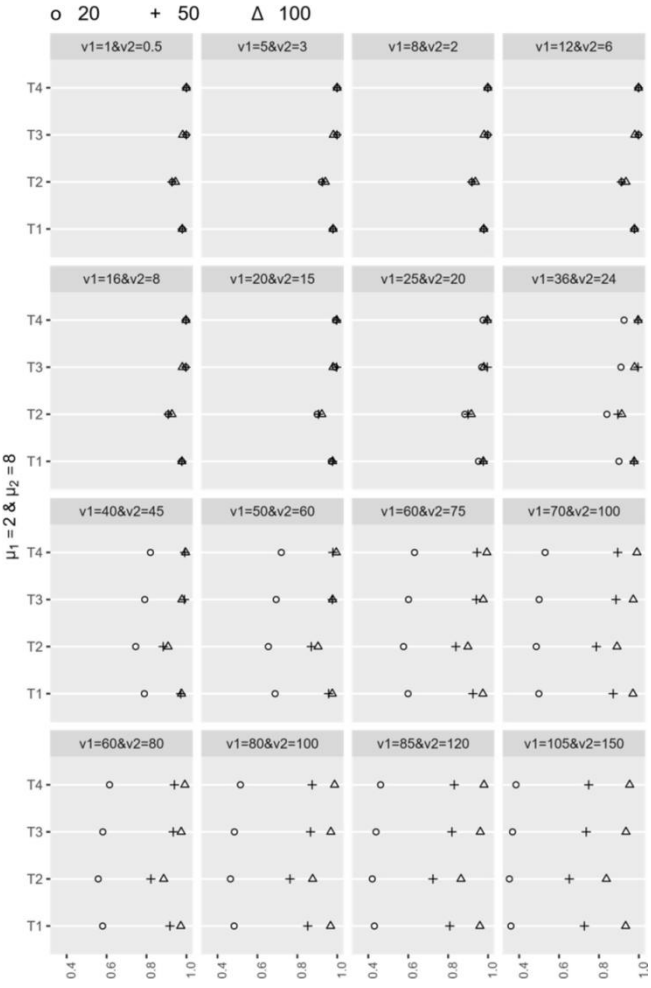

Figure 2. The estimated power of the four tests

By Figures. 3, 4, and 5 we can get the simplified overview about the estimated power of the test for the four tests when the data is balanced. These figures corresponding to the values of Tables (4, 5 and 6) respectively.

In Figure. 3, the power of the test for tests T1, T3, and T4 is very close in almost cases. While, the power of the test for T2 is the lowest generally.

In Figure. 4, the power of the test for tests T2 is the lowest one in all cases. And the gap of the power between test T2 and other tests became clearer.

Also, the same result that shows in Figure. 5. The power of the test for T2 is the lowest. But, the power of the test for T4 is the best power. Then, the power of the test for T1 in the second level. Also, the power of the test for T3 test is closer to the power for T1.

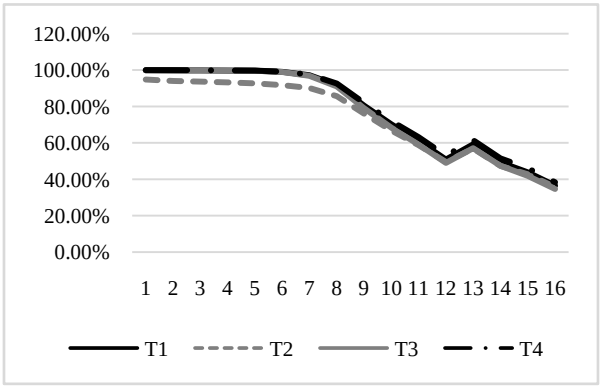

Figure 3. The estimated power of the four tests

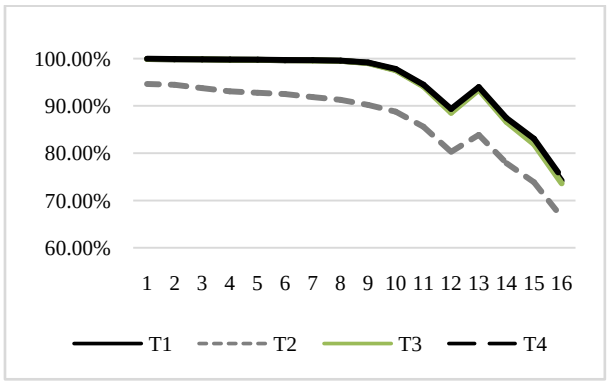

Figure 4. The estimated power of the four tests

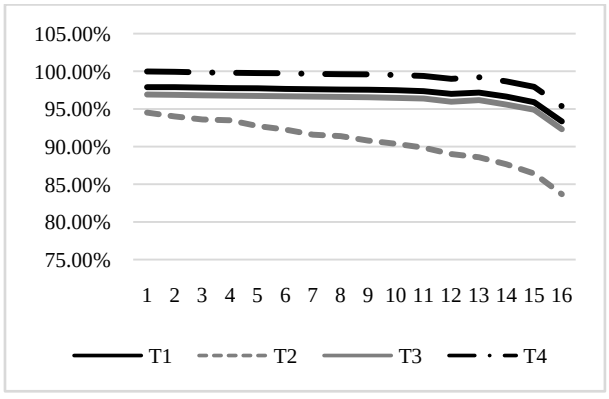

Figure 5. The estimated power of the four tests

Table 7 shows the estimated type-I error probabilities for the four tests when the sample sizes are not equal (unbalanced data) at a nominal level $\alpha = 0.05$, $\mu_k = 2$ under

different variances.

Table 7. The Probability of Type-I Error for The Four Tests Under Different Variances, Different Sample Sizes (Unbalanced Data) and $\mu_k = 2$

Var (1)	Var (2)	n	m	T1	T2	T3	T4
1	0.5	5	4	0.037	0.180	0.025	0.067
5	3	8	5	0.044	0.127	0.033	0.064
8	2	10	8	0.042	0.090	0.034	0.055
12	6	12	10	0.042	0.083	0.036	0.054
16	8	25	8	0.043	0.085	0.037	0.053
20	15	40	25	0.041	0.064	0.039	0.046
25	20	60	15	0.044	0.075	0.042	0.053
36	24	85	10	0.043	0.082	0.038	0.053
40	45	100	30	0.042	0.064	0.042	0.047
50	60	130	20	0.043	0.070	0.041	0.051
60	75	170	50	0.047	0.066	0.047	0.050
70	100	200	100	0.040	0.055	0.042	0.043
60	80	150	40	0.044	0.063	0.043	0.048
80	100	175	15	0.046	0.080	0.042	0.055
85	120	200	10	0.041	0.085	0.039	0.053
105	150	225	5	0.044	0.134	0.044	0.067

Figure. 6 represents the estimated type-I error probabilities (transformed by $\log_{10}$) for the four tests under different variances for unbalanced data (sample sizes are different) at $\mu_k = 2$. This figure corresponds to values that are listed in Table 7.

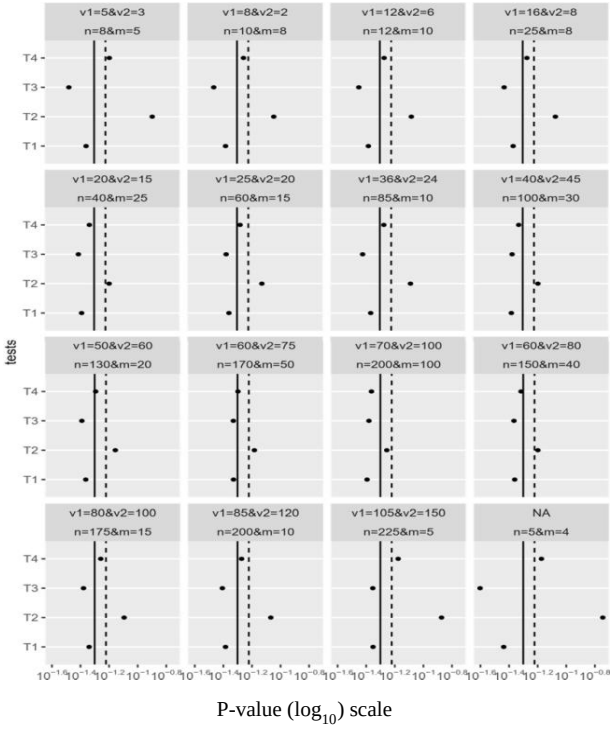

Figure 6. The estimated probabilities of type-I error for the four tests

Figure. 6 shows that the estimated type-I error probabilities for test T4 overestimate type-I error when the sample sizes and variances are small but it become closer to the nominal level when increasing sample sizes and variances. Also, type-I error probabilities for tests T1 and T3 are so far from a nominal level (0.05). However, they become closer to 0.05 when sample sizes and variances are increasing. Finally, the estimated type-I error probabilities for test T2 overestimate in most cases and so far from 0.05 in other cases. In both cases, it shows non-acceptable size so, the power of test T2 would not be reliable.

Table 8 represents the power of the four tests when the sample sizes are not equal and ($\mu_1 = 2$, $\mu_2 = 8$) under different variances.

Table 8. The Power of The Test for The Four Tests Under Different Variances, Different Sample Sizes and ($\mu_1 = 2$, $\mu_2 = 8$)

Var(1)	Var(2)	n	m	T1	T2	T3	T4
1	0.5	5	4	87.15%	85.32%	90.44%	91.57%
5	3	8	5	87.12%	85.31%	90.37%	91.55%
8	2	10	8	86.98%	85.30%	90.07%	91.47%
12	6	12	10	86.65%	85.08%	89.71%	91.21%
16	8	25	8	87.14%	85.31%	90.43%	91.56%
20	15	40	25	86.43%	85.00%	89.50%	91.01%
25	20	60	15	86.36%	84.77%	89.43%	90.85%
36	24	85	10	78.21%	76.76%	79.79%	83.80%
40	45	100	30	87.08%	85.25%	90.37%	91.50%
50	60	130	20	86.39%	84.64%	89.63%	90.80%
60	75	170	50	86.15%	84.45%	89.28%	90.57%
70	100	200	100	76.48%	74.65%	78.41%	81.16%
60	80	150	40	84.38%	82.82%	88.35%	88.77%
80	100	175	15	48.59%	47.26%	48.94%	53.65%
85	120	200	10	28.99%	28.22%	29.16%	33.91%
105	150	225	5	11.95%	10.05%	12.42%	16.93%

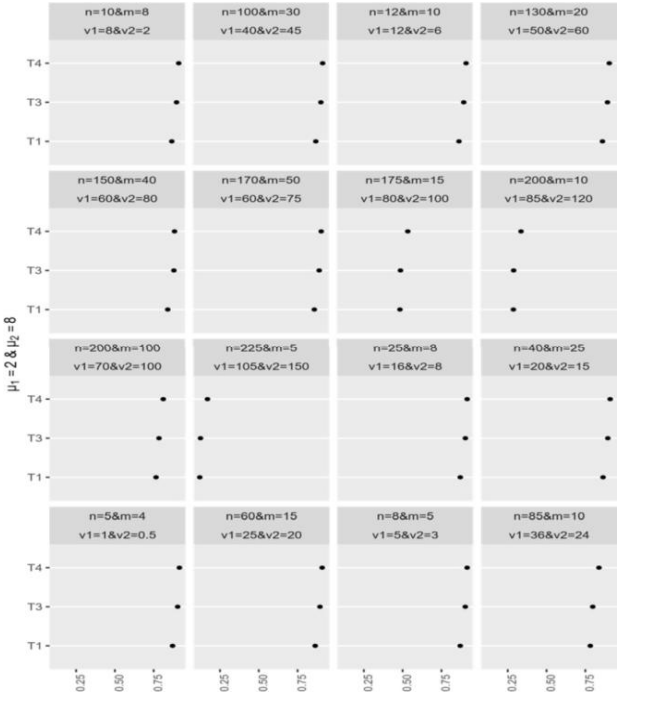

Figure 7. The estimated power of the three tests

Figure. 7 represents the power of the test for three tests only. This figure corresponds to the numerical values founded in Table 8. In this figure, test T2 has been deleted from the power comparison, because the size of this test is not acceptable sizes.

Figure. 7 shows that the power of T4 test is the best power among the three tests in this comparison. The power for T3 is better than T1. But, the power of all tests reached the lowest level when one sample size is small while, variances values and the gap between them are large.

In Figure. 8 we can show the overview about the power of the test for three tests only (T1, T3 and T4) when the data is unbalanced. This figure corresponds to the numerical values in Table 8. This figure shows that the power of T4 test is the best power among tests in this comparison. Also, the power for T3 is better than T1.

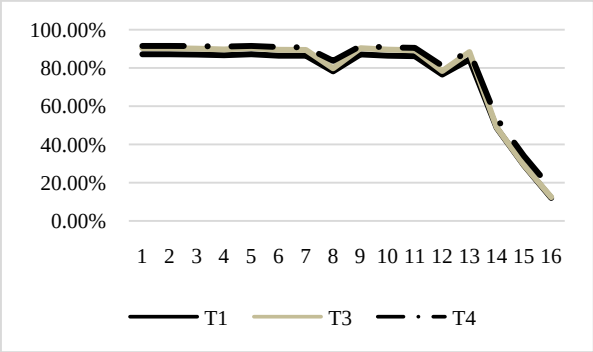

Figure 8. The estimated power of the three tests

5. Summary and Conclusions

In this paper, two tests (T3, and T4) were proposed to solve the B-F problem. These tests depended on the test statistic which was introduced by Behrens (1929) using the variance estimators. These estimators based on the method that was provided by Chen et al. (2022) that based on Fisher's fiducial argument to estimate the variances of the sample means $v(\bar{x}_1)$,$v(\bar{x}_2)$ respectively. For each suggested test statistic, we needed to get the degrees of freedom (f) and the constant (C) to approximate the test statistic to t- distribution as we shown in Welch approximation. Then, we derived the formula for degree of the freedom (f) and the constant (C) for each suggested solution to approximate their distributions to t-distribution as we shown in Welch approximation and Fenstad test.

Monte Carlo simulation was used to evaluate the performance of the proposed tests (T3, and T4) and the other tests such as (Welch test, and Fenstad test) under several scenarios. The simulation study was conducted to compare the sizes (the estimated type-I error probabilities) and the powers of these tests. This simulation study was based on three factors (i) sample sizes (balanced or unbalanced), (ii) values of population variances, and (iii) the gap between population variances. The main findings of the simulation study can be summarized in the following:

1) The estimated type-I error probabilities for tests T4, and T1 (Welch test) are closer to a nominal level 0.05 when the sample sizes are equal (balances data) especially when the sample sizes are large as shown in Figure. 1.
2) In most cases, test T2 has overestimated probability of type-I error. Therefore, test T2 cannot be recommended for testing the differences between two population means generally.
3) When sample sizes and variances are small, the estimated type-1 error for test T3 is so far from a nominal level. But it becomes closer to nominal level with increasing the sample sizes and variances. So, we can recommend test statistic T3 to deal with B-F problem when the sample sizes and variances are large.
4) When the data is unbalanced, the type-I error probabilities for tests T3, and T1 are so far from the nominal level. However, they become closer to nominal level when sample sizes and variances are increasing. This result about test T1 agrees with Chen's mention in his study in 2022 (Welch test is applicable for large sample sizes only).
5) The estimated type-I error probabilities for test T4 is overestimated when the sample sizes and variances are small but it becomes closer to the nominal level when increasing sample sizes and variances.
6) When the data is balanced, the power for test T4 is better than the power for test T1 when the variances are small regardless of the sample sizes. But, the power for test T1 is better than the power of T3 slightly.
7) In general, the power for tests T1, T3, and T4 are decreasing when the gap between variances gets larger and sample sizes are equal, but still T4 with the highest power.
8) When sample sizes are unequal, the powers for T4 and T3 are better than the power of T1. Where the power for test T4 is the best power between all tests in this comparison.
9) Generally, when the sample sizes are unequal, one sample size is much smaller than the other, and the gap of the variances are large, the power of all tests reached to the lowest level.

Finally, we conclude that the proposed tests (T3 and T4) can be recommended as alternative new solutions to the B-F problem especially, when sample sizes are large. That is because, the power of test T3 and test T4 are better than or close to the power of test T1 (Welch test).

REFERENCES

[1] Aoki, S. "Effect Sizes of the Differences between Means without Assuming Variance Quality and between a Mean and a Constant." *Heliyon* 6 (2020).

[2] Behrens, W V. "Ein Beitrag Zur Fehlerberechung beiwenigen Beobachtungen." *Landwirtsch.* (Jahrbucher) 68 (1929): 807-837.

[3] Best, D. J., and J. C. Rayner. " Welch's Approximate Solution for the Behrens-grimes Problem." *Technometrics* 29 (1987): 205-2010.

[4] Chen, CH., Yilin Li, K. Liang, and J. Du. "A Test for the Behrens-Fisher Problem Based on the Method of Variance Estimates Recovery." *Communication in Statistic- Theory Methods* 51 (2022).

[5] Cochran, W. G. "Approximation Significance Levels of the Behrens-Fisher Test." *Biometrics* 20 (1964): 191-195.

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

[7] Fisher, R. A. "The Comparison of Samples with Possibly Unequal Variances." *Annals of Eugenics* 9 (1939): 174-180.

[8] Grimes, B. A., and W. T. Federer. "Comparison of Means from Populations with Unequal Variances." (Biometrics Unit Series, Cornell University, Ithaca, new york) 1982.

[9] Ibrahim, I. H. "On The Behrens-Fisher Problem and The Bootstrabe Solution An Alternative Approch." *Journal of the faculty of commerch for scientific research, faculty of commece, Alexandria university* XXXVII (2000).

[10] Kim, S. H., and A. S. Cohen. "On the Behrens-Fisher Problem: A Review." *Journal of Educational and Behavioral Statistics* 23 (1998): 356-377.

[11] Ozkip, E., B. Yazici, and A. Sezer. "A simulation Study on Tests for the Behrens- Fisher Problem." *Turkiye Klinikleri J Biostat* 6 (2014): 59- 66.

[12] Paul, S. R., D. J. Best, and J. C. W. Rayner. "Comment on Best and Rayner (1987)." *Technometrics* 34 (1992): 249-250.

[13] Paul, S. R., Y. G. Wang, and I. Ullah. "A Review of the Behrens-Fisher Problem and Some of Its Analogs: Does the Same Size Fit All?" *Revstat Statistical Journa* 4 (2019): 563-597.

[14] Scariano, S. M., and B. S. "A Four Moment Solution to The Behrens- Fisher Problem." (Texas Tech. university) 1981.

[15] Welch, B. L. "The Significance of the Difference between Two Means when the Population Variances are Unequal." *Biometrika* 29 (1938): 350-362.