American Journal of Mathematics and Statistics 2023, 13(2): 76-83
DOI: 10.5923/j.ajms.20231302.02

New Modified Test for 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 is the problem of testing equality of two population means using two independent samples when the quotient of the population variances is unknown. Ibrahim et al. (2023) introduced two modified tests that are based on the method that was provided by Chen et al. (2022). In this paper, a new modified test is proposed which depends on Fisher's fiducial argument to estimate the variances of the sample means. A comprehensive simulation study is designed for balanced and unbalanced samples with different sizes and various ranges of variances. The simulation study shows that the power of the suggested modified test is outperform Welch test and the two tests of Ibrahim et al. (2023) especially, for large sample sizes and wide range of variances of two independent samples with balanced or unbalanced sample sizes.

Keywords Balanced data, Behrens-Fisher problem, Fisher's fiducial argument, New modified test, Two modified tests, Unbalanced data, Welch test

1. Introduction

The Behrens-Fisher (B-F) problem is the problem of testing the hypothesis of equality of the means of two normal populations using two independent samples when the population variances are unknown or with possibly unequal variances [18] [8] [3] [1].

Several solutions have been developed and these solutions are divided into parametric solutions and non-parametric solutions. Behrens (1929) introduced a parametric test that was the first parametric solution of B-F problem and confirmed by Fisher (1939) but the estimated type I error of this test is frequently smaller than the nominal level [10] [8] [3] [1]. Many solutions proposed for B-F problem, such as the Welch test as an approximation solution that introduced by Welch (1938). Welch approximation is the popular approximation solution for the B-F problem. Also, several approximation solutions have been proposed, such as Cochran Approximation (1964), Fenstad (1983), Wald test, which proposed by Best and Rayner (1987) [18] [6] [5] [3]. The latest parametric solutions have been proposed by Ibrahim et al. (2023), Chen et al. (2022), and Hong et al. (2022).

In this study, we propose new modified test using the same technique used in [10] by making modifications to the assumed value of the random variable, as shown in section 3.

The new test is based on Fisher's fiducial argument to estimate the variances of the sample means, as the method that was suggested by Chen et al. (2022). This study aims to compare a new modified test with Welch test, which is widely used to deal with the B-F problem, and two tests that were introduced in [10]. This comparison was made by using a comprehensive Monte Carlo simulation with different scenarios and factors to assess the size and the power of these tests. This simulation study consists of three factors: (i) sample sizes (ii) balanced or unbalanced data, (iii) various wide ranges of populations' variances values to assess the impact of the gap between population variances. This study proceeds as follow: Section 2 introduces available solutions to the B-F problem. Then, Section 3 presents the suggested solutions. Section 4 demonstrates the simulation study. Finally, Section 5 shows the conclusion of the study.

* Corresponding author:
ghadataha@commerce.helwan.edu.eg (Ghada Taha)
Received: Nov. 2, 2023; Accepted: Nov. 13, 2023; Published: Nov. 29, 2023
Published online at http://journal.sapub.org/ajms

2. Available Solutions to the B-F Problem

Many solutions have been introduced to the B-F problem. These solutions can be classified into exact and approximated solutions. In this paper, we focused on the approximation solutions. These solutions, such as (i) the Welch test (T1), (ii) the two tests that have been introduced by Ibrahim et al. (2023) (T2 and T3).

For testing the equality between two normal populations means when the variances are unknown or unequal based on two independent samples: the first sample $x_1,\ldots, x_n$ from $N(\mu_1,\sigma_1^2)$ and $y_1,\ldots, y_m$ from $N(\mu_2,\sigma_2^2)$; where $-\infty<\mu_k<\infty$, $0<\sigma_k^2<\infty$, and k =1, 2. The null and alternative hypotheses are:

$$H_0: \mu_1=\mu_2 \text{ (or } \mu_1-\mu_2=0) \text{ vs. } H_1: \mu_1>\mu_2 \text{(or } \mu_1-\mu_2>0).$$

Initially, we define some statistics as:

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

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

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

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

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

Therefore

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

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

Then,

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

Therefore,

$$E(s_1^2)=\sigma_1^2,\ E(s_2^2)=\sigma_2^2 \quad (7)$$

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

i) Welch test (T1): This is a popular test of B-F problem, which proposed by Welch (1938). This test is considered as the standard solution to testing the equality between two population means from an independent normal population with unequal variances [6] [4] [3]. The Welch statistic T1:

$$T1=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}} \quad (8)$$

Where, T1 approximated by t-distribution with degrees of freedom ($f_{(1)}$) as:

$$f_{(1)}=\frac{\left(\frac{S_1^2}{n}+\frac{S_2^2}{m}\right)^2}{\left(\frac{\left(\frac{S_1^2}{n}\right)^2}{(n-1)}\right)+\left(\frac{\left(\frac{S_1^2}{m}\right)^2}{(m-1)}\right)} \quad (9)$$

ii) Two tests of Ibrahim et al. (2023) (T2 and T3):

$$T2=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{(n-1)S_1^2}{n^2}+\frac{(m-1)S_2^2}{m^2}}}=T.GH.1 \quad (10)$$

$$T3=\frac{(\bar{x}_1-x_2)}{\sqrt{\frac{(n-1)s_1^2}{n(n-2)}+\frac{(m-1)s_2^2}{m(m-2)}}}=T.GH.2 \quad (11)$$

Where, T2, T3 approximated to t-distribution with degrees of freedom ($f_{(2)}$), ($f_{(3)}$) and constant ($C_{(2)}$), ($C_{(3)}$), respectively. The degrees of freedom are:

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

$$f_{(3)}=\frac{\left(\frac{(n-1)s_1^2}{n(n-2)}+\frac{(m-1)s_2^2}{m(m-2)}\right)^2}{\left(\frac{\left(\frac{(n-1)s_1^2}{n(n-2)}\right)^2}{(n-1)}\right)+\left(\frac{\left(\frac{(m-1)s_2^2}{m(m-2)}\right)^2}{(m-1)}\right)} \quad (13)$$

The constants are:

$$C_{(3)}=\frac{\frac{S_1^2}{n}+\frac{S_2^2}{m}}{\frac{(n-1)s_1^2}{n^2}+\frac{(m-1)s_2^2}{m^2}} \quad (14)$$

$$C_{(4)}=\frac{\frac{S_1^2}{n}+\frac{S_2^2}{m}}{\frac{(n-1)s_1^2}{n(n-2)}+\frac{(m-1)s_2^2}{m(m-2)}} \quad (15)$$

3. Suggested Solution to the B-F Problem

In this paper, we suggest a new modified test as an alternative to t-test when the B-F problem occurs. This test is based on the method that was proposed by Chen et al. (2022) with some modifications which depends on Fisher's fiducial argument to estimate the variances of the sample means. The test statistic, degrees of freedom, and constant could be derived as follows:

Let T be the test statistic:

$$T=\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}} \quad (16)$$

$$T=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{v(\bar{x})+v(\bar{y})}} \quad (17)$$

When H_0 is true, the test statistic can be rewritten as:

$$T=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}}=\frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{v(\bar{x})+v(\bar{y})}} \quad (18)$$

This test statistic could be approximated by the student t-distribution as a Welch approximation. Where: $T\sim c\, t_f$.

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

$$v(\bar{x})+v(\bar{y})=\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m} \quad (19)$$

Where, σ_k^2 is often unknown when the B-F problem occurs, we can use the variance estimate $\hat{\sigma}_k^2$ instead of σ_k^2. Then, equation (19) can be written as:

$$\widehat{v(\bar{x})}+\widehat{v(\bar{y})}=\frac{\hat{\sigma}_1^2}{n}+\frac{\hat{\sigma}_2^2}{m} \quad (20)$$

We can estimate σ_j^2 by using the following relationships [4]:

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

Let:

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

Where, U_k is a random variable that follows a chi-square distribution with k degrees of freedom.

Therefore:

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

To estimate σ_j^2, let some values for U_k to get the values of $\hat{\sigma}_k^2$ corresponding to them. Various values for U_k will lead to various values of $\hat{\sigma}_k^2$.

As shown in [10], we can get the degrees if freedom f and constant c as the following formulas:

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

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

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

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

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

In [10], Chen et al. (2022) U_k is replaced by (n-3), which is the maximum value of the probability density function. If we use the same method to estimate the variances at (U_k=n-1), we will get the Welch statistic. Where, (n-1) is the mean of the chi-square distribution. If we replace U_k with (n-) T.GH.1 as it has been shown in [10]. Also, we can get T.GH.2 when we replace U_k by (n-2). Thus, in this paper, we proposed a new test statistic by replacing U_k by (n-4) and follow the same method to estimated variances as follows:

Assuming that U_1=n-4 and U_2=m-4, In this case, we can get the estimated variances by replacing the variables (U_1, U_2) with (n-4, m-4), respectively. Therefore, we can reformulate equation (22) as:

$$\hat{\sigma}_1^2 = \frac{(n-1)s_1^2}{n-4},\ \hat{\sigma}_2^2 = \frac{(m-1)s_2^2}{m-4} \quad (26)$$

$$\widehat{v(\bar{x})} = \frac{(n-1)s_1^2}{n(n-4)},\ \widehat{v(\bar{y})} = \frac{(m-1)s_2^2}{m(m-4)} \quad (27)$$

Then, we can use $\widehat{v(\bar{x})}$, $\widehat{v(\bar{y})}$ in equation (18) to get the test statistic T4 as:

$$T4 = \frac{(\bar{x} - \bar{y})}{\sqrt{\frac{(n-1)s_1^2}{n(n-4)} + \frac{(m-1)s_2^2}{m(m-4)}}} = T.New \quad (28)$$

We can approximate this test statistic T4 to t-distribution with $f_{(4)}$ degrees of freedom and constant $C_{(4)}$ as:

$$T4 \sim C_{(4)} t_{f_{(4)}} \quad (29)$$

And we can get $f_{(4)}$ and $C_{(4)}$ by using the formulas in equations (23) and (24), respectively as the following:

$$f_{(4)} = \frac{\left(\frac{(n-1)s_1^2}{n(n-4)} + \frac{(m-1)s_2^2}{m(m-4)}\right)^2}{\left(\frac{\left(\frac{(n-1)s_1^2}{n(n-4)}\right)^2}{(n-1)}\right) + \left(\frac{\left(\frac{(m-1)s_2^2}{m(m-4)}\right)^2}{(m-1)}\right)} \quad (30)$$

$$\therefore f_{(4)} = \frac{\left(\frac{(n-1)s_1^2}{n(n-4)} + \frac{(m-1)s_2^2}{m(m-4)}\right)^2}{\frac{(n-1)S_1^4}{n^2(n-4)} + \frac{(m-1)S_2^4}{m^2(m-4)}} \quad (31)$$

$$C_{(4)} = \frac{\frac{s_1^2}{n} + \frac{s_2^2}{m}}{\frac{(n-1)s_1^2}{n(n-4)} + \frac{(m-1)s_2^2}{m(m-4)}} \quad (32)$$

4. Simulation Study

The simulation studies were designed to compare four tests (T1- T.GH.1- T.GH.2- T.New) by using the Monte Carlo method and applying these simulations on R- package as shown in the following steps:

1- The data for samples were generated randomly from the normal populations at various configurations of the factors (μ_1, μ_2, n, m, Var(1), Var(2)).
2- For each simulation, the sample means ($\bar{x}$, $\bar{y}$) were estimated.
3- For each simulation, the variances were estimated ($\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$).
4- The test statistics were calculated for the four tests (T1 - T.GH.1- T.GH.2- T.New).
5- The probability of Type-I error or (size of the test) was calculated for the four tests (T1 - T.GH.1- T.GH.2- T.New).
6- The power of the test were calculated for the four tests (T1 - T.GH.1- T.GH.2- T.New).

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

(1) (Welch test (T1),
(2) The first proposed test in [10] is represented by (T2 or T.GH.1).
(3) The second proposed test in [10] is represented by (T3 or T.GH.2).
(4) The new suggested test in this paper (T.New).

These simulation studies were based on three factors: (i) sample sizes (ii) balanced or unbalanced data, (iii) various wide ranges of populations' variances values to assess the impact of the gap between population variances. The simulation studies were conducted on different scenarios to compare the probability of Type-I error or size and power of each test under various configurations. These studies are conducted on samples generated from the normal populations with different means and different variances in two scenarios:

Case 1: Balanced data.

Case 2: Unbalanced data.

These simulation studies were based on 10000 generated samples at a significance level α = 0.05, the samples generated from the normal distribution at $\mu = 2$ and different variances. Where Var(1) and Var(2) are the variances of the first and second population, respectively. The estimated Type-I error probabilities for the four tests (T1 - T.GH.1- T.GH.2- T.New) for balanced samples at (n, m = 20, 50, and 100) are shown in tables (1, 2, and 3).

Table 1. The Probability of Type-I Error for The Four Tests (T1 - T.GH.1- T.GH.2- T.New) Under Different Variances, $\mu_k = 2$ and n = m = 20

Var(1)	Var(2)	T1	T.GH.1	T.GH.2	T.New
1	0.5	0.0495	0.0427	0.0548	0.05512
5	3	0.0492	0.0436	0.055	0.0572
8	2	0.0478	0.0447	0.0556	0.05632
12	6	0.0516	0.0464	0.0574	0.0576
16	8	0.0484	0.0421	0.055	0.05624
20	15	0.0484	0.044	0.0537	0.05488
25	20	0.0529	0.0471	0.0606	0.06
36	24	0.048	0.0423	0.0522	0.0544
40	45	0.0494	0.0446	0.0556	0.05616
50	60	0.0489	0.0428	0.0559	0.05656
60	75	0.0472	0.041	0.052	0.05456
70	100	0.0508	0.0435	0.0568	0.05736
60	80	0.0481	0.0426	0.0545	0.05704
80	100	0.049	0.0436	0.0564	0.0564
85	120	0.0444	0.0392	0.0513	0.052
105	150	0.0511	0.0458	0.0567	0.0576

Table 2. The Probability of Type-I Error for The Four Tests (T1 - T.GH.1- T.GH.2- T.New) Under Different Variances, $\mu_k = 2$ and n =m = 50

Var(1)	Var(2)	T1	T.GH.1	T.GH.2	T. New
1	0.5	0.0483	0.04719	0.052155	0.05385
5	3	0.0493	0.04785	0.05225	0.0513
8	2	0.0439	0.04356	0.047975	0.048375
12	6	0.0485	0.04884	0.05282	0.05205
16	8	0.0464	0.04598	0.05111	0.051225
20	15	0.051	0.05104	0.052345	0.052125
25	20	0.0492	0.04928	0.052535	0.052425
36	24	0.0445	0.04169	0.048735	0.050475
40	45	0.0516	0.05126	0.054625	0.0534
50	60	0.0494	0.04895	0.053485	0.054
60	75	0.0494	0.04917	0.051585	0.052875
70	100	0.0472	0.04653	0.051395	0.051525
60	80	0.0493	0.04785	0.053295	0.0534
80	100	0.0519	0.05269	0.056335	0.05565
85	120	0.0501	0.04994	0.054625	0.05295
105	150	0.0496	0.04829	0.05282	0.05265

As we showed in the previous table, the probability of Type-I error for Welch test (T1) is closer to the significance level 5% than other tests (T.New (T4), T.GH.2 (T3)). Also, these tests are acceptable sizes in all configurations at 5%. But, the probability of Type-I error for T.GH.1 (T2) is far from 5% compared with other tests.

By increasing the sample sizes to 50 and 100, as we showed in tables 2 and 3, the probability of Type-I error for all tests is acceptable. The estimated Type-I error for Welch test is the closest to the significance level 5%.

Table 3. The Probability of Type-I Error for The Four Tests (T1 - T.GH.1- T.GH.2- T.New) Under Different Variances, $\mu_k = 2$ and n= m = 100

Var(1)	Var(2)	T1	T.GH.1	T.GH.2	T.new
1	0.5	0.0505	0.0499	0.0517	0.0541
5	3	0.0541	0.0529	0.0552	0.0581
8	2	0.0492	0.0484	0.0505	0.0529
12	6	0.0522	0.0512	0.0532	0.0554
16	8	0.0515	0.0505	0.0527	0.0552
20	15	0.0491	0.0477	0.0497	0.0511
25	20	0.0493	0.0479	0.0497	0.0519
36	24	0.0502	0.0486	0.0514	0.0542
40	45	0.0525	0.0506	0.0535	0.0565
50	60	0.0515	0.05	0.0524	0.0549
60	75	0.0474	0.0463	0.0482	0.0495
70	100	0.0482	0.0469	0.0499	0.0527
60	80	0.0556	0.0543	0.0569	0.0588
80	100	0.0484	0.0476	0.0495	0.0524
85	120	0.0507	0.0495	0.0516	0.0542
105	150	0.0505	0.0495	0.0515	0.0537

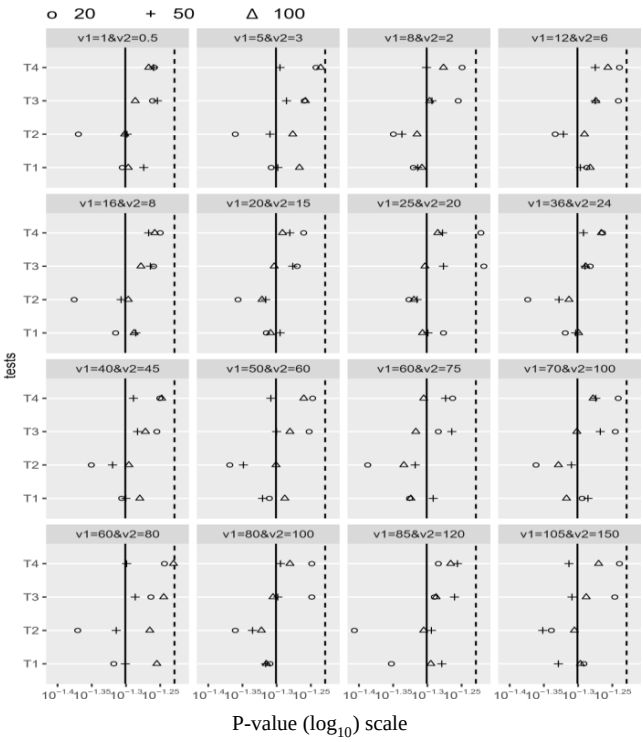

Figure 1. The estimated probabilities of Type-I error for the four tests

The estimated Type-I error probabilities (transformed by log10) for the four tests (T1 - T.GH.1- T.GH.2- T. New) in

Tables (1, 2 and 3) can be represented graphically in Figure 1. In Figures 1,` and 3 as we can see two vertical lines and different symbols. These two lines, represent the solid and broken lines equivalent to 0.05 and 0.06, respectively. The different symbols represent the different sample sizes (n, m = 20, 50 and 100) as shown in Figures 1, 2, 3, and 4.

Table 4. The Power of the Test for the Four Tests (T1 - T.GH.1- T.GH.2- T. New) Under Different Variances, $\mu_1 = 2$, $\mu_2 = 8$ and n=m=20

Var(1)	Var(2)	T1	T.GH.1	T.GH.2	T. new
1	0.5	99.59%	99.61%	99.63%	99.69%
5	3	99.01%	99.12%	99.23%	99.51%
8	2	98.91%	99.03%	99.06%	99.11%
12	6	98.87%	98.89%	98.98%	98.87%
16	8	97.96%	94.93%	94.96%	94.97%
20	15	97.26%	94.22%	94.37%	94.57%
25	20	95.41%	92.19%	92.85%	93.44%
36	24	90.28%	86.71%	88.40%	89.75%
40	45	78.81%	75.05%	77.80%	80.52%
50	60	68.45%	64.51%	68.01%	71.31%
60	75	59.84%	56.00%	60.11%	63.93%
70	100	50.62%	47.47%	51.17%	55.07%
60	80	57.91%	54.25%	58.07%	61.75%
80	100	49.74%	46.37%	50.28%	54.37%
85	120	42.85%	39.63%	43.42%	47.76%
105	150	37.69%	34.83%	38.15%	42.14%

Table 5. The Power of the Test for the Four Tests (T1 - T.GH.1- T.GH.2- T.New) Under Different Variances, $\mu_1 = 2$, $\mu_2 = 8$ and n= m= 50

v1	v2	T1	T.GH.1	T.GH.2	T.New
1	0.5	99.95%	99.97%	99.98%	100.00%
5	3	99.88%	99.91%	99.93%	99.96%
8	2	99.87%	99.87%	99.90%	99.92%
12	6	99.78%	99.83%	99.89%	99.90%
16	8	99.75%	99.77%	99.84%	99.85%
20	15	99.66%	99.73%	99.79%	99.80%
25	20	99.64%	99.66%	99.71%	99.74%
36	24	97.98%	97.98%	97.98%	97.98%
40	45	97.61%	97.59%	97.65%	97.69%
50	60	96.00%	95.90%	96.11%	96.33%
60	75	93.23%	93.06%	93.38%	93.74%
70	100	88.33%	88.09%	88.68%	89.29%
60	80	92.67%	92.48%	92.85%	93.19%
80	100	86.45%	86.00%	86.87%	87.65%
85	120	81.68%	81.27%	82.21%	83.06%
105	150	73.45%	72.75%	74.09%	75.45%

As we will show in Figure 1, the probability of Type-I error for tests T1 (Welch test), T. GH.2, and (T.New) is closed to a nominal probability of 0.05 in all combinations (acceptable size). But the probably of Type-I error for test T.GH.1 is far from the significance level 5% when the sample sizes are small and become close to 5% when the sample sizes are increasing. But the size of this test becomes closer to 0.05 when the sample size and the value of variances are increasing.

Table 6. The Power of the Test for the Four Tests (T1 - T.GH.1- T.GH.2- T.New) Under Different Variances, $\mu_1 = 2$, $\mu_2 = 8$ and n=m= 100

Var(1)	Var(2)	T1	T.GH.1	T.GH.2	T.New
1	0.5	96.93%	95.96%	99.00%	99.42%
5	3	96.90%	95.92%	98.97%	99.37%
8	2	96.87%	95.85%	98.89%	99.31%
12	6	96.83%	95.82%	98.83%	99.27%
16	8	96.78%	95.78%	98.80%	99.23%
20	15	96.70%	95.75%	98.77%	99.19%
25	20	96.67%	95.70%	98.67%	99.15%
36	24	96.64%	95.65%	98.62%	99.11%
40	45	96.59%	95.60%	98.63%	99.05%
50	60	96.55%	95.52%	98.54%	98.97%
60	75	96.39%	96.38%	97.39%	98.39%
70	100	96.14%	96.13%	97.14%	98.15%
60	80	96.34%	96.34%	97.34%	98.36%
80	100	95.88%	95.81%	96.91%	97.92%
85	120	95.14%	95.08%	96.20%	97.25%
105	150	93.00%	92.94%	94.05%	95.23%

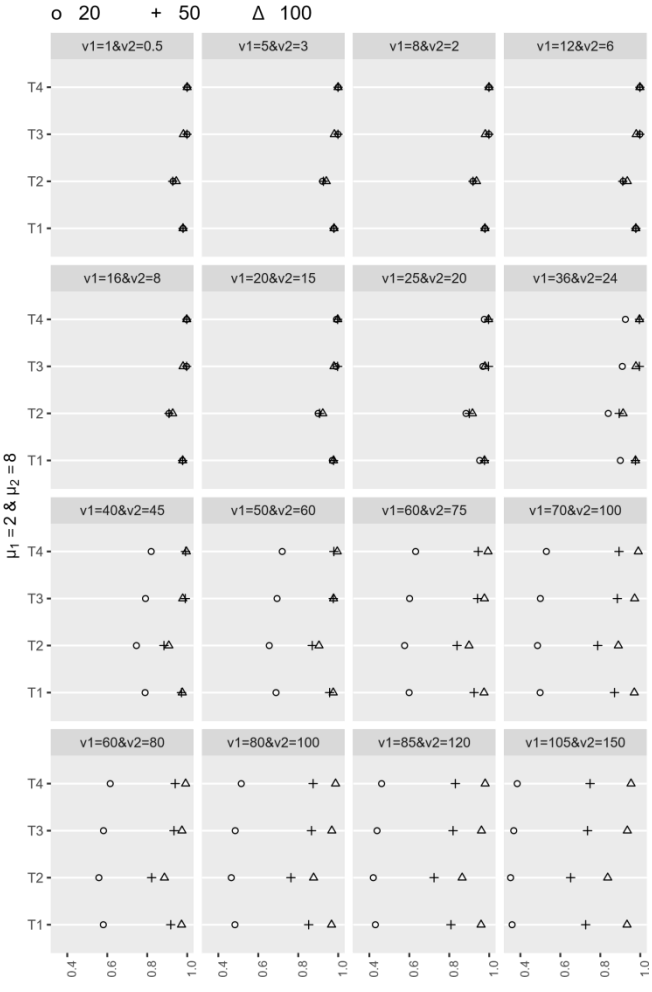

Figure 2. The estimated power of the four tests

The results of the power of the test for the four tests (T1 - T.GH.1- T.GH.2- T.New) can be seen in Tables (4, 5, and 6). These powers were calculated for balanced data that were generated at (μ_1= 2, μ_2 = 8) under different variances.

We can represent the results in Tables (4, 5, and 6) graphically, as we shown in Figure 2.

The power of the test for test T4 (T.New) is the best in all cases, reaching to 100% when the sample sizes are large (n, m >30) and variances are small. The power for test T3 (T. GH.2) is better than the power for T1 (Welch test) in most cases, regardless of the variance values at the large sample size. Also, the power for test T3 (T. GH.2) is better than the power for T1 (Welch test) in most cases. Generally, the power of the test for the four tests is decreasing with increasing the values of the variances and gap between the variance values.

In Figures (3, 4, and 5) the values of the power for the four tests in Tables (4, 5 and 6) can be represented graphically in the simplified graphs to give an overview of opinion about the estimated power of the tests when the data is balanced.

As we can see in Figures (3, and 4) the power of the test for tests T1 (Welch test), T2 (T.GH.1), T3 (T.GH.2), and T4 (T.New) is very close in most cases because the difference between them is small.

In Figure. 5, the power of the test for T4 (T. New) is the best. Also, the power of the test for T1 is better than other tests (T2 (T.GH.1), T3 (T.GH.2)).

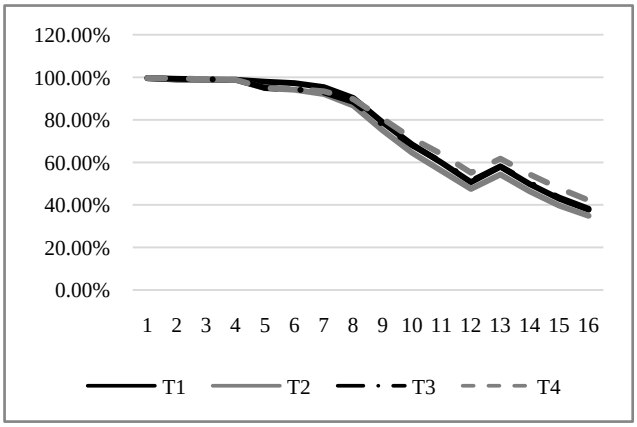

Figure 3. The estimated power of the four tests

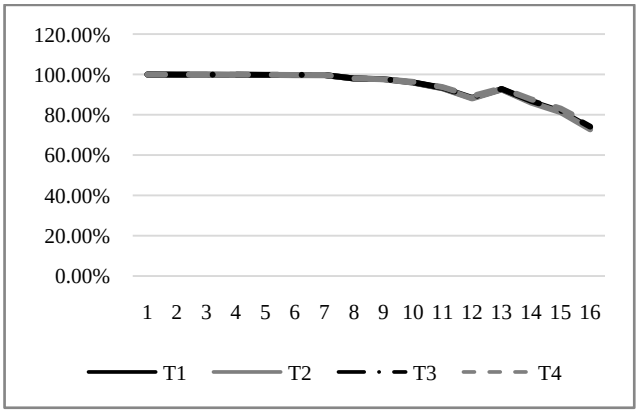

Figure 4. The estimated power of the four tests

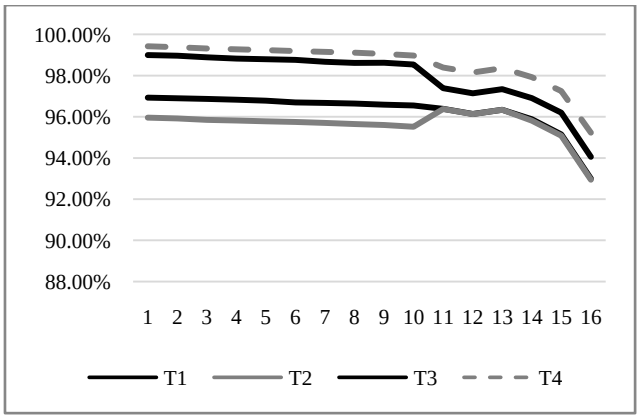

Figure 5. The estimated power of the four tests

In Table 7, the estimated Type-I error probabilities for the four tests (T1 - T.GH.1- T.GH.2- T.New) can be shown when the sample sizes are unbalanced data at a significance level $\alpha = 0.05$, and $\mu_k = 2$ under different variances.

Table 7. The Probability of Type-I Error for The Four Tests (T1 - T.GH.1- T.GH.2- T.New) Under Different Variances, Different Sample Sizes (Unbalanced Data) and $\mu_k = 2$

Var(1)	Var(2)	n	m	T1	T.GH.1	T.GH.2	T.New
1	0.5	20	12	0.0483	0.0427	0.0564	0.0579
5	3	25	15	0.0485	0.0421	0.0545	0.0593
8	2	35	20	0.0485	0.0452	0.0531	0.0508
12	6	55	25	0.0498	0.0463	0.0529	0.0506
16	8	50	25	0.0487	0.0451	0.0528	0.0501
20	15	80	35	0.0493	0.0476	0.0524	0.0473
25	20	100	30	0.0532	0.0499	0.0555	0.0512
36	24	120	40	0.0484	0.046	0.0505	0.0524
40	45	100	25	0.0506	0.0469	0.055	0.0556
50	60	120	30	0.0484	0.0443	0.0516	0.0496
60	75	150	45	0.0497	0.0478	0.0525	0.0557
70	100	180	40	0.0481	0.046	0.0513	0.0495
60	80	150	15	0.0501	0.0436	0.0569	0.0446
80	100	200	20	0.0491	0.044	0.0546	0.0478
85	120	225	25	0.0505	0.047	0.0549	0.056
105	150	250	30	0.0502	0.0459	0.0537	0.0513

In Figure 6, the estimated Type-I error probabilities in Table 7. after transformed by $\log_{10}$ for the four tests (T1 - T.GH.1- T.GH.2- T.New) can be represented graphically for unbalanced data at $\mu_k = 2$ under different variances.

Figure 6 shows that the estimated Type-I error probabilities for the test T2 is far from the significance level 5% especially when the sample sizes and variances are small, and become closer to this significance level when both sample sizes and variances are increasing.

Also, this figure showed that the estimated Type-I error probabilities for other tests are acceptable.

Also, in Table 8 the power of the four tests can be shown when the sample sizes are unbalanced at (μ_1 = 2, μ_2 = 8) under different variances.

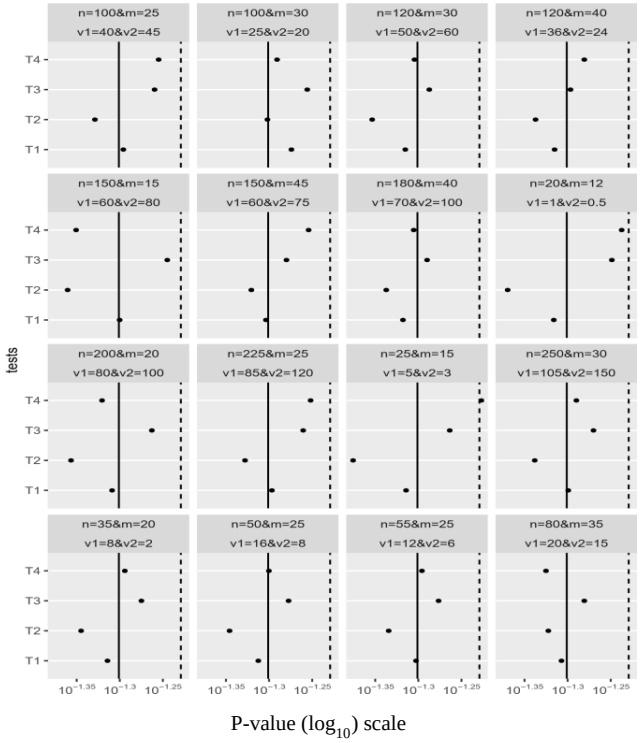

Figure 6. The estimated probabilities of Type-I error for the four tests

Table 8. The Power of the Test for the Four Tests (T1 - T.GH.1- T.GH.2- T.New) Under Different Variances, Different Sample Sizes and ($\mu_1 = 2$, $\mu_2 = 8$)

Var (1)	Var (2)	n	m	T1	T.GH.1	T.GH.2	T. New
1	0.5	20	12	89.99%	91.99%	93.07%	94.45%
5	3	25	15	86.67%	85.55%	87.68%	89.99%
8	2	35	20	93.22%	92.77%	93.61%	94.57%
12	6	55	25	83.37%	82.68%	84.35%	85.92%
16	8	50	25	68.75%	67.70%	69.88%	72.09%
20	15	80	35	67.16%	66.46%	67.92%	69.60%
25	20	100	30	53.88%	53.03%	55.06%	57.28%
36	24	120	40	54.66%	53.90%	55.33%	56.77%
40	45	100	25	26.21%	25.12%	27.24%	29.59%
50	60	120	30	24.01%	23.19%	24.97%	26.63%
60	75	150	45	28.85%	28.22%	29.44%	30.68%
70	100	180	40	21.10%	20.53%	21.84%	23.12%
60	80	150	15	12.86%	11.89%	14.36%	17.51%
80	100	200	20	13.78%	12.89%	14.68%	17.27%
85	120	225	25	14.52%	13.73%	15.24%	16.98%
105	150	250	30	13.53%	12.89%	14.09%	15.61%

Figure 7 shows the power of the test of the four tests obtained in Table 8. This figure shows that the power of T.New is the best. Also, we can conclude that the power of the test of T3 (T.GH.2) is better than the power of the test of T1 (Welch test) in all cases. When the variances increase, the power of all tests will decrease for unbalanced data.

In Figure 8, the power of the test of the four tests (T1, T2, T3 and T4) that are shown in Table 8 can be represented graphically for unbalanced data.

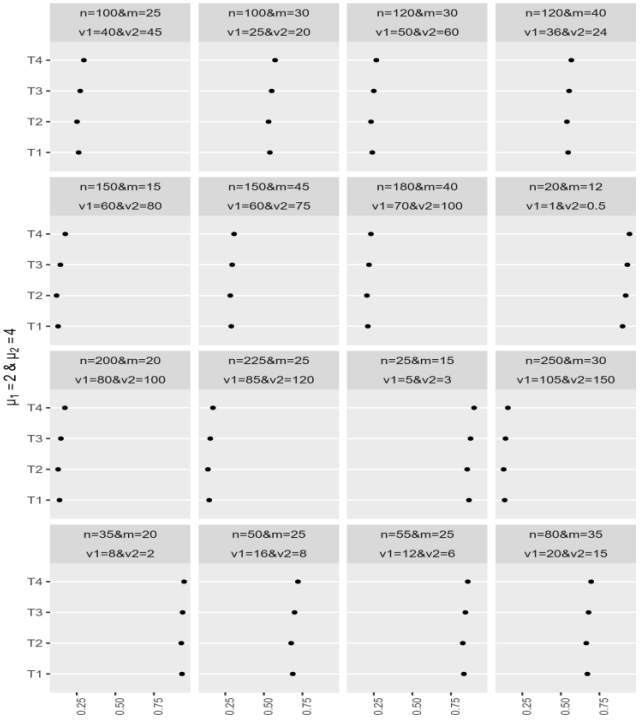

Figure 7. The estimated power of the four tests

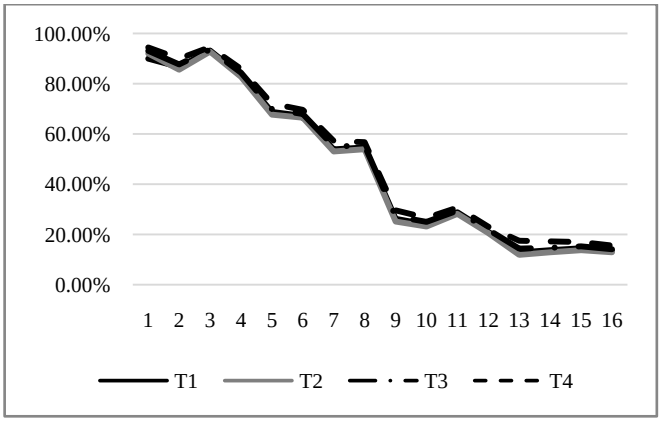

Figure 8. The estimated power of the four tests

Figure 8 shows that the new modified test (T.New) is outperform other compared tests for the estimated power of the test.

5. Summary and Conclusions

In this paper, we suggested a new modified test (T.New) as an alternative for t-test when the homogeneity assumption is violated. This new modified test is based on the method that was proposed by Chen et al. (2022) with some modifications which depends on Fisher's fiducial argument to estimate the variances of the sample means. The degrees of freedom (f) and the constant (C) for the new modified test are derived to approximate the new modified test statistic to t- distribution as shown in Welch approximation. A comprehensive simulation study with different factors and scenarios has been conducted to evaluate the performance of

the new modified test comparing with Welch test and the two other tests that have been introduced by Ibrahim et al. (2023). This comparison was based on the size of the tests and the power of the tests. The main statistical findings can be summarized in the following:

1) The probability of Type-I error probabilities for tests T1 (Welch test), T.GH.2, and T.new (the new suggested test) are acceptable when the data is balanced and small sample sizes at $\mu = 2$ with different variances.
2) The probability of Type-I error for all tests is acceptable when increasing the sample sizes to 50 and the data are balanced at $\mu = 2$ and variances different.
3) In most cases, the power of the test for the new suggested test T4 is the best when the data is balanced at $\mu = 2$ under different variances.
4) In most cases, the power for test T3 (T. GH.2) is better than the power for T1 (Welch test) for the balanced data at $\mu = 2$ under different variances.
5) Generally, the power of the test for the four tests is very close in most cases, as previously shown in Figures 3 and 4. These powers are decreasing when increasing the values of variances and gap between the values of variances for the balanced sample sizes.
6) When the data is unbalanced, the estimated Type-I error probabilities for test T2 are far from the significance level 5% for the small sample sizes and variances.
7) By increasing the sample sizes and variances, the estimated Type-I error probabilities for test T2 become closer to this significance level 5%.
8) When the data is unbalanced, the estimated Type-I error probabilities for tests (T1, T3, and T4) are acceptable.
9) The power of the test for T4 (T.New) is the best for unbalanced data in all cases studied.
10) Also, the power of the test for T3 (T.GH.2) is better than the power of the test for T1 (Welch test) in all cases studied.
11) When the variances increase, the power of all tests will decrease for the unbalanced data.

Finally, we conclude that the suggested test T4 (T.New) has the best power when compared to the other tests and can be recommended to use as an alternative test for t-test when the homogeneity assumption is violated.

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] Hong, s., A. Gelhoc, and J. Park. "An Exact and Near-Exact Distribution Approach to the Behrens–Fisher Problem." *Mathematics* 10 (2022).

[10] Ibrahim, I. H., GH. Taha, and M. Sadek "Parametric Solutions to the Behrens-Fisher Problem." *American Journal of Mathimatics and Statistics* 13 (2023): 60-68

[11] 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).

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

[13] Larsen, R. J., and M. L. Marx. *An Introduction to Mathematical Statistics and Its Application.* 5. Pearson Education, 2011.

[14] 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.

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

[16] 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.

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

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