# ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA

## DEPARTMENT OF STATISTICS

## A Review work on Comparison of different Approaches related to Behrens - Fisher testing Problem through simulation



**NAME**: ANIRBAN GHOSH

**ROLL NO**: 439

**REGISTRATION NUMBER**: A01-1112-0835-19

**SEMESTER**: 6

**SESSION**: 2019-2022

**SUPERVISOR**: PROF. DEBJIT SENGUPTA

### Declaration

*I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.*

*Anirban Ghosh*

_____

**ANIRBAN GHOSH**

# ABSTRACT

The Behrens-Fisher problem arises when one seeks to make inferences about the means of two normal populations without assuming that the variances are equal. This paper consists of a review work of fundamental concepts and applications used to address the Behrens-Fisher problem under fiducial approach, welch approach and Banerjee's approach. The difficulty with the Behrens-Fisher problem is that exact solutions are not available satisfactorily because nuisance parameters are present. A table for tests of significance is also included.

# INTRODUCTION

A frequently encountered problem in applied statistics is testing the difference between two population means. The Behrens – Fisher problem arises when one seeks to make inferences about the means of two normal populations without assuming either that the variances are equal or the ratio of variances is known. Under such conditions, Neyman-Pearson (1928,1933) sampling theory may provide a different solution from those available via either Bayesian theory (e.g, Jeffreys, 1940) or Fisher's (1935,1939) fiducial theory (Kendall & Stuart, 1979; Lehmann, 1993). Although a number of methods have been proposed for the Behrens-Fisher problem beginning with Behrens (1929) and Fisher (1935), no definitive solutions exist (Robinson, 1982). For reasonably large sample sizes, differences between various extent solutions are generally much less than between these solutions and use of Student's t test. When sample sizes are small, however the three theories may yield different solutions.

In the context of Bayesian and fiducial theories, several sets of tables (Fisher &Healy, 1956; Fisher &Yates, 1957; Issacs, Christ Novick &Jackson, 1974; Lindley & Scott, 1984; Sukhatme, Thawani, Pendharkar, & Natu, 1951) have been presented for the Behrens – Fisher problem. No tables are available, however for directional hypothesis testing for even numbers of degrees of freedom and small sample sizes.

This paper presents a review work of the Behrens – Fisher problem, focusing on fundamental concepts and applications rather than theoretical and philosophical considerations. It consists of three parametric tests or approaches to the Behrens- Fisher problem. We will also discuss here the method of approximations of Behrens- Fisher distribution.

## **Some important definitions**

**Statistical Test:** As the term suggests, one wishes to decide whether or not some hypothesis That has been formulated is correct or not. To collect information regarding the population (distribution), we do experimentation and collect data. Then, based on the data collected, we make decision. Using some rule, whether to reject or accept the hypothesis.

The choices here lie between two decisions: Rejection or Acceptance of the hypothesis. Then A decision rule for such a problem is called a **Test of the hypothesis**.

**Definition:** A statistical test of a statistical hypothesis H is a rule or procedure of deciding whether to accept H or reject H, based on a random sample from the population.

**Testing Problem:**

Consider a population distribution with distribution function F(x ;θ) , $\theta \in \Omega$

Consider two hypotheses of interest H₁: $\theta \in \Omega_0$ and H₂: $\theta \in \Omega_1$ , where $\Omega_0 \cap \Omega_1 = \phi$ and $\Omega_0 \cup \Omega_1 = \Omega^* \subseteq \Omega$ .

here $\Omega^*$ is known as parameter space of interest. Let $(x_1, x_2, x_3, \dots \dots, x_n)$ be an observed sample from the population distribution. The data provides information regarding population. Based on the data and a testing rule to test or examine whether the data supports H₁ or H₂ and if the data supports H₁ then we accept H₁

**Test and critical region:**

Let $\mathfrak{X}$ be the collection of all possible samples of size n, be the sample space or potential data set.

Consider a test: To test $H_0: \theta \leq 17$ against $H_1: \theta > 17$ ; for the population distribution N(θ,5²). Let x=(x₁,x₂,……,xₙ) be a sample of size n from N(θ,5²)

Reject H₀ iff $\bar{x} > 17 + \frac{5}{\sqrt{n}}$ : Testing rule

A test procedure assigns to each possible sample, one of two decisions: Reject H₀ or accept H₀; Thereby it divides the sample space $\mathfrak{X}$ into two complementary parts W and $\mathfrak{X}$ −W=Wᶜ, such that if x falls into W, then H₀ is rejected; otherwise, it is accepted. The set W is called the **Rejection Region** or **Critical Region** of the test.

**Performance of A test:**

| Decision<br>True State | Reject $H_0$ | Accept $H_0$ |
|---|---|---|
| $H_0$ is true | Wrong (Type I error) | Correct |
| $H_0$ is false | Correct | Wrong (Type II error) |

While performing a test, one may arrive at a correct decision or may commit one of the two wrong decisions or the two errors:

    i)       Rejecting $H_0$ when it is true [Type I error]

    ii)      Accepting $H_0$ when it is false [Type II error]

The two types of error probabilities are given by

(1) The probability of rejecting $H_0$ when it is true $=P[x \in w | \theta]$ when $\theta \in \Omega_0$ is called

    **Probability of Type I error**.

(2) The probability of accepting $H_0$ when it is false $= P[x \in w^c | \theta]$ when $\theta \in \Omega_1$

$$= 1\text{-}P[x \in w | \theta] \text{ when } \theta \in \Omega_1$$

    is called **Probability of Type II error**.

**Construction of the test:** The usual procedure of finding a test $H_0$ against $H_1$ is to restrict that error probability which is more serious i.e the probability of type-I-error and then to minimize the probability of type-II-error as far as possible.

Hence one selects a number $\alpha \in [0,1]$ and impose restriction on tests on tests given by critical region W,

$P[X \in W | \theta] \le \alpha$ , for all $\theta \in \Omega_0$

**Level of significance of a test:** The quantity $\alpha$ is called the level of significance of testing problem of testing $H_0$ vs $H_1$. The choice of level of significance of course depends on the experimenter himself. If he thinks that the false rejection of $H_0$ (type -I-error) is more serious error, then he will rather take a small value of α, say 0.01, 0.001 etc. On the other hand, if he thinks that this error is not so serious, he will not mind taking a value of α high as α=0.05,0.1

**Size of a test:** For a test given by the critical region W, the number Supremum of $P[x \in W | \theta]$

when $\theta \in \Omega_0$ is called the size of the test W.

The size of a test gives the maximum possible probability of committing Type I error

**Remark:** For a testing problem $H_0: \theta \in \Omega_0 \ vs \ H_1: \theta \in \Omega_1$ , with a chosen level of significance α, there are several tests and a test W may have size > α, then the test W will not be considered for construction of a good test.

**Power of a test:** The probability of rejecting H₀ when it is false i.e P[x∈ $w|\theta$], evaluated at $\theta = \theta_1 \in \Omega_1$, is called the power of test W, for testing $H_0: \theta \in \Omega_0$, against the alternating value $\theta = \theta_1$

[Clearly the power of a test is the probability of taking one of the two correct decisions arising in the testing problem.]

# Foundation Of the Behrens-Fisher Problem

We motivate the discussion of the Behrens-Fisher problem with an example taken from Marascuilo and Serlin (1988). In this problem patients in a mental health clinic were given one of two initial treatments: $n_1$ =4, number of patients received a film treatment and $n_2 = 3$, number of patients received an interview treatment. Researchers wanted to know whether the patients in the film treatment and interview treatment differed in the number of times they returned to the clinic for the subsequent treatment. The following are the data for this problem: x₁= $(x_{11}, x_{12}, \ldots, x_{1n_1}) = (8,10,12,15); \overline{x_1} = 11.25$ , s₁² = 8.91667, x₂= $(x_{21}, x_{22}, \ldots \ldots, x_{2n_2}) =$ (1,7,11); $\overline{x_2} = 6.33333$ , s₂² =25.33333, where the sample mean and sample variance for i = 1,2 are defined as

$$\overline{x_\iota} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \ \ldots\ldots\ldots\ldots\ldots (1)$$

and

$$s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_\iota})\ldots\ldots\ldots\ldots (2)$$

We assume that the two independent samples, x₁ⱼ and x₂ⱼ were drawn from two normal distributions having means $\mu_1$ and $\mu_2$ and variances $\sigma_1^2$ and $\sigma_2^2$ respectively. With the assumptions of equal variances that is $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , the population variance is estimated by

the pooled sample variance, $s^2 = 15.48333$, using

$$s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-1}\dots\dots\dots\dots (3)$$

The sufficient statistics for $\mu_1, \mu_2,$ and $\sigma^2$ are $\overline{x_1}, \overline{x_2}$ and $s^2$. Note also that $\overline{x_2} - \overline{x_1}$ has a normal distribution with mean $\delta = \mu_2 - \mu_1$ and variance $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2$. The student's t pivotal statistic with $n_1 + n_2 - 2$ degrees of freedom is

$$t = \frac{\delta - (\overline{x_2}-\overline{x_1})}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s^2}} \sim t_{n_1+n_2-2} \dots\dots\dots\dots (4)$$

If we denote $t_{\frac{\alpha}{2}}(v)$ as the value for which

$$\Pr\{t > t_{\frac{\alpha}{2}}(v)\} = \frac{\alpha}{2} \dots\dots\dots (5)$$

And also denote

$$\alpha = \Pr\{t < -t_{\frac{\alpha}{2}}(v)\} + \Pr\{t > t_{\frac{\alpha}{2}}(v)\}\dots\dots\dots (6)$$

Then the $100(1-\alpha)\%$ confidence interval for $\delta$ is

$$\overline{x_2} - \overline{x_1} \pm \frac{t_{\alpha}}{2}(n_1 + n_2 - 2)\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s^2} \dots\dots\dots\dots (7)$$

For the null hypothesis is $H_0 : \delta = 0$, we have $t_{obs} = 1.63599$ with 5 degrees of freedom . The resulting two tailed p value is 0.16277 with a 95% confidence interval of [-12.64209,2.80875] or $-12.64209 \leq \delta \leq 2.80875$. The result is that the difference between the two means is not significant at the 0.05 level.

When it is not reasonable to assume $\sigma_1^2 = \sigma_2^2$ ; neither a pivotal statistic nor an exact confidence interval procedure exists. One simple way to solve this problem, however is to use a proxy

$$t* = \frac{\delta - (\overline{x_2}-\overline{x_1})}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_{[\min(v_1,v_2)]}\dots\dots\dots (8)$$

where $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$. Restate in terms of a confidence interval, we have

$$\overline{x_2} - \overline{x_1} \pm t_{\frac{\alpha}{2}}[\min(v_1, v_2)]\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \ldots\ldots\ldots\ldots (9)$$

Note that, in this case (i.e t* test), the actual confidence is greater than $100(1 - \alpha)\%$

The marascuilo and serlin (1988) data yield t* = 1.50493 with 2 degrees of freedom, the

Resulting two tailed p value is 0.27127. The 95% confidence interval is [-18.97365,9.14031].

When $n_1 = n_2 = n$. the value obtained from equation 4 is the same as that from equation 8. In

this special case we can use equations 4 and 7 to test the null hypothesis and to obtain the

100(1-α) % confidence interval, respectively. However as noted in Hsu (1938) and Robinson

(1976), the Type I error probability might be greater than the specified nominal level unless

the equality of two variance is satisfied.

The above t* solution is a simple approach to the Behrens-Fisher problem. The other approach

based on test statistic t* are derived from fiducial theory. We first present a solution from

fiducial theory.

## Fisher's Fiducial Approach

Fisher (1935) proposed a statistical method for obtaining a probability distribution of a

parameter from observed data called a fiducial probability distribution. Recall that we assume

the two sets of observations are random samples drawn from independent normal distributions.

The quantities $\overline{x_i}$ and $s_i^2$ are jointly sufficient for $\mu_i$ and $\sigma_i^2$ having independent sampling

distributions $N(\mu_i, \frac{\sigma_i^2}{n_i})$ and $\left(\frac{\sigma_i^2}{v_i}\right)\chi_{v_i}^2$ , respectively for i = 1,2. Consequently we can define

$$t_i = \frac{\mu_i - \overline{x_i}}{\sqrt{\frac{s_i^2}{n_i}}} \sim t_{n_i-1} \ldots\ldots\ldots\ldots (10)$$

By logical inversion

$$\mu_i = \overline{x_i} + t_i\sqrt{\frac{s_i^2}{n_i}} \ldots\ldots\ldots\ldots (11)$$

and $\delta = \mu_2 - \mu_1$ is given by

$$\delta = \overline{x_2} - \overline{x_1} + t_2\sqrt{\frac{s_2^2}{n_2}} - t_1\sqrt{\frac{s_1^2}{n_1}} \ldots\ldots\ldots\ldots (12)$$
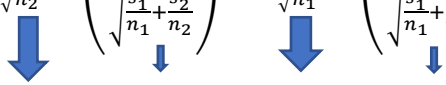
The fiducial distribution of $\delta$ can be used to make fiducial inferences about $\delta$ and to set the

fiducial intervals. Instead of obtaining the distribution of $\delta$, for the purpose of tabulation,

however, Fisher (1935,1939) chose the statistic

$$T = \frac{\delta - (\overline{x_2} - \overline{x_1})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = t_2 cos\theta - t_1 sin\theta \; \ldots\ldots\ldots (13)$$

Where θ is taken in the first quadrant; the algebraic details are as follows

$$T = \frac{(\mu_2 - \mu_1) - (\overline{x_2} - \overline{x_1})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\mu_2 - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} - \frac{\mu_1 - \overline{x_1}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{\mu_2 - \overline{x_2}}{\frac{s_2}{\sqrt{n_2}}} \cdot \left( \frac{\frac{s_2}{\sqrt{n_2}}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right) - \frac{\mu_1 - \overline{x_1}}{\frac{s_1}{\sqrt{n_1}}} \cdot \left( \frac{\frac{s_1}{\sqrt{n_1}}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right) (14)$$

This is t₂    This is $cos\theta$    This is t₁    This is $sin\theta$

And $tan\theta = \dfrac{\frac{s_1}{\sqrt{n_1}}}{\frac{s_2}{\sqrt{n_2}}}$    ……….. (15)

Note that $T$ is the same quantity as t* but for ease of presentation in the context of both fiducial and Bayesian theories we use $T$. The distribution of $T$ is the Behrens-Fisher distribution and is defined by the three parameters $v_1$, $v_2$ and θ, $T = T(v_1, v_2, \theta)$. The distribution can be seen as a mixture of two t distributions, $t_{v_1}$ and $t_{v_2}$

The Behrens – Fisher distribution is actually the conditional distribution of the quantity (14) given the values of the quantities labeled as $cos\theta$ and $sin\theta$

If we denote $T_{\frac{\alpha}{2}}(v_1, v_2, \theta)$ as the value for which

$$\Pr\{T > T_{\frac{\alpha}{2}}(v_1, v_2, \theta)\} = \frac{\alpha}{2} \; \ldots\ldots\ldots (16)$$

Then the $100(1 - \alpha)\%$ fiducial interval is

$$\overline{x_2} - \overline{x_1} \pm T_{\frac{\alpha}{2}}(v_1, v_2, \theta) \cdot \left( \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

Using the fiducial limits the interval estimates of $\mu_2 - \mu_1$ was obtained.

However, Fisher approximated the distribution of $T$ by ignoring the random variation of the relative sizes of the standard deviations, $\dfrac{\frac{s_1}{\sqrt{n_1}}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Fisher's solution provoked controversy because it did not have the property that the hypothesis

of equal means would be <u>rejected with probability α</u> if the means were in fact equal. Many other methods of treating the problem have been proposed since, and the effect on the resulting confidence intervals have been investigated.

# Welch's approximate t solution

Mean comparison is the central theme of many classical statistical procedures. The well-known independent-sample t-test is often used to test the equality of two means from independent populations with equal variances, whereas Welch's t-test is generally preferred when the variances are not equal.

The testing problem is to test $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 \neq 0$

B.L Welch used the test statistic which is given by

$$T_1 = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)}$$

Under H$_0$, T$_1$ can be written as

$$T_1 = \frac{(\overline{x_1} - \overline{x_2})}{\left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)}$$

Where the distribution of T$_1$ is approximated by Student's t distribution with degrees of freedom $\nu$ under the null hypothesis where $\nu$ is determined by the equation

$$\frac{1}{\nu} = \frac{w^2}{n_1 - 1} + \frac{(1-w)^2}{n_2 - 1}$$

Where w is defined as

$$w = \frac{\frac{s_1^2}{n_1}}{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

And here $\nu$ is a random variable depending upon the sample size n$_1$ and n$_2$

As the sample sizes go to infinity then T$_1$ approximately follows Standard normal Distribution i.e N (0,1)

[ Actually, in general this $\nu$ is unknown to us and it is approximated as

$$\frac{1}{\nu} = \frac{w^2}{n_1 - 1} + \frac{(1-w)^2}{n_2 - 1}$$

Where w $= \dfrac{\frac{\sigma_1^2}{n_1}}{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}$ where $\sigma_1^2$ and $\sigma_2^2$ are population variances which are unknown to us.

$\nu$ can also be approximated as

$$\nu \approx \frac{(\gamma_1+\gamma_2)^2}{\frac{\gamma_1^2}{n_1-1}+\frac{\gamma_2^2}{n_2-1}} \quad \text{where } \gamma_i = \frac{\sigma_i^2}{n_i}$$

Under the null hypothesis of equal expectations, $\mu_1 = \mu_2$, the distribution of the Behrens–Fisher statistic $T$, which also depends on the variance ratio $\sigma_1{}^2/\sigma_2{}^2$, could now be approximated by **Student's t distribution** with these $\nu$ degrees of freedom. But this $\nu$ contains the population variances $\sigma_i{}^2$, and these are unknown. The following estimate only replaces the population variances by the sample variances:

$$\hat{\nu} \approx \frac{(g_1+g_2)^2}{\frac{g_1^2}{n_1-1}+\frac{g_2^2}{n_2-1}} \quad \text{where } g_i = \frac{s_i^2}{n_i}$$

This $\hat{\nu}$ is a random variable. A t distribution with a random number of degrees of freedom does not exist. Nevertheless, the Behrens–Fisher $T$ can be compared with a corresponding quantile of **Student's t distribution** with these estimated numbers of degrees of freedom, $\hat{\nu}$ , which is generally non-integer. In this way, the boundary between acceptance and rejection region of the test statistic $T$ is calculated based on the empirical variances $s_i{}^2$, in a way that is a smooth function of these.]

This method also does not give exactly the nominal rate, but is generally not too far off. However, if the population variances are equal, or if the samples are rather small and the population variances can be assumed to be approximately equal, it is more accurate to use **Student's t-test**.

## Banerjee's Approach

Saibal Kumar Banerjee proposed another approach on finding the confidence interval for linear functions of means of k populations when there are no valid assumptions of equality of population variances.

It is shown that given k samples of $n_i$ units from populations $N(\mu_i , \sigma_i^2)$ ; $(i = 1,2, \dots k)$

A confidence interval for any linear function $\sum_{i=1}^{k} c_i \mu_i$ of population means with confidence coefficient not less than any pre-assigned probability (1-α) is possible in terms of sample estimates of population means and variance and tabulated values of student's t- table.

### Derivation:

Given k samples of $n_i$ units from k normal populations $N(\mu_i, \sigma_i^2)$ with sample estimates of Population means and variances $\overline{x_i}$ and $s_i^2$ $(i = 1,2, \dots k)$ a confidence interval for any linear function of the population means with approximate confidence coefficient is possible. Briefly the method is indicated for the case of two samples (i.e k = 2).

Let P denote the probability of the event

$$(\overline{x_1} + \overline{x_2} - \mu_1 - \mu_2)^2 \le \frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2} \quad \dots \dots (1)$$

Where $t_1$ and $t_2$ correspond to t-values of student's t table satisfying the relation

$$\frac{1}{\sqrt{v_i}} \frac{1}{B\left(\frac{v_i}{2},\frac{1}{2}\right)} \int_{-t_i}^{t_i} \left(1 + \frac{t^2}{v_i}\right)^{-\frac{v_i+1}{2}} dt = 1 - \alpha \quad \dots \dots (2)$$

$$\text{Where } v_i = n_i - 1 \; ; i = 1,2$$

For fixed $s_1^2$ and $s_2^2$

$$\left(\frac{\overline{x_1} + \overline{x_2} - \mu_1 - \mu_2}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)^2$$

is distributed $\chi^2$ variate with 1 degrees of freedom.

Here the critical region or rejection region is

$$w: \left\{ (\overline{x_1} + \overline{x_2} - \mu_1 - \mu_2)^2 \ge \frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2} \right\}$$

Where $\mu_1 + \mu_2 = \xi_0$ under the null hypothesis

It can be shown by suitable mathematical manipulation that,

$$\text{Prob} \left\{ (\overline{x_1} + \overline{x_2} - \mu_1 - \mu_2)^2 \le \frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2} \right\} \ge 1 - \alpha \quad \dots \dots (3)$$

So that

$$\text{Prob} \left\{ \overline{x_1} + \overline{x_2} - \sqrt{\sum_{i=1}^{2} \frac{t_i^2 s_i^2}{n_i}} \le \mu_1 + \mu_2 \le \overline{x_1} + \overline{x_2} + \sqrt{\sum_{i=1}^{2} \frac{t_i^2 s_i^2}{n_i}} \right\} \ge 1 - \alpha \quad \dots \dots (4)$$

So, the $100(1 - \alpha)\%$ confidence interval for $\mu_1 + \mu_2$ is given by

$$\left( \overline{x_1} + \overline{x_2} - \sqrt{\sum_{i=1}^{2} \frac{t_i^2 s_i^2}{n_i}} \ , \ \overline{x_1} + \overline{x_2} + \sqrt{\sum_{i=1}^{2} \frac{t_i^2 s_i^2}{n_i}} \right)$$

Also, it can be readily shown that if $c_1$ and $c_2$ are known constants then

$$\text{Prob } \{ \sum_{i=1}^{2} c_i \overline{x_\iota} - \sqrt{\sum_{i=1}^{2} \frac{t_i^2 s_i^2 c_i^2}{n_i}} \leq \sum_{i=1}^{2} c_i \mu_i \leq \sum_{i=1}^{2} c_i \overline{x_\iota} + \sqrt{\sum_{i=1}^{2} \frac{t_i^2 s_i^2 c_i^2}{n_i}} \} \geq 1 - \alpha \ \dots\dots (5)$$

Further extending to k populations it can be shown that

$$\text{Prob } \{ \sum_{i=1}^{k} c_i \overline{x_\iota} - \sqrt{\sum_{i=1}^{k} \frac{t_i^2 s_i^2 c_i^2}{n_i}} \leq \sum_{i=1}^{k} c_i \mu_i \leq \sum_{i=1}^{k} c_i \overline{x_\iota} + \sqrt{\sum_{i=1}^{k} \frac{t_i^2 s_i^2 c_i^2}{n_i}} \} \geq 1 - \alpha \ \dots\dots.. (6)$$

For the case of two populations if $c_1 = 1$ and $c_2 = -1$ in (5) the following relation is established

$$\text{Prob } \{ |\overline{x_1} - \overline{x_2} - (\mu_1 - \mu_2)| \leq \sqrt{\frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2}} \} \geq 1 - \alpha$$

which is the two means problem. Cochran and Cox had suggested an approximate result for this case which is given by

$$|\overline{x_1} - \overline{x_2} - (\mu_1 - \mu_2)| \leq \frac{\left( \frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2} \right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Since

$$\left( \frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2} \right)^2 \leq \left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right) \left( \frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2} \right)$$

If $n_1 \neq n_2$ this result gives a stronger result than previous result as the $100(1-\alpha)$ % confidence interval has become wider in the previous case suggested by Saibal Banerjee, so the interval estimation is less precise.

To the best of the knowledge of the author he has not seen the result suggested as approximation proved in any published literature nor the approximate nature of the result spelt out. The approximate nature of the present result is to affect that for all values of $(\sigma_1^2, \sigma_2^2)$ or $(\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2)$ the true value will be covered with probability not less than pre-assigned probability $(1-\alpha)$

## Other Approaches related to Behrens-Fisher Problem

The Behrens-Fisher problem has been well known since the early 1930's.

Many approaches other than these were proposed by many statisticians. Among those approaches Scheffe's approach was most well-known and satisfactory. One reason for it's popularity is that there is no exact solution satisfying the classical criteria for good test. For example, Fisher, Welch, Aspin, Cochran and Cox, Qin and Jing have all suggested different solutions. 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. Tsui and Tang apply the distributional property of the generalized p-value for the Behrens-Fisher problem to multiple testing. Kim and Cohen propose a review of fundamental concepts and applications to address the Behrens-Fisher problem under fiducial, Bayesian and frequentist approaches. Singh, Saxena, Srivastava proposed a new test using jackknife methodology. Dong considers the empirical likelihood approach for this problem. Chang and Pal revisit BF problem and apply a newly developed Computational Approach Test (CAT)

## Simulation Study:

**Objective:**

It is desirable to carry out a test in such a manner which keep both type of error probabilities at a minimal level. Unfortunately for a sample of fixed size n both error probabilities can not be controlled simultaneously.

[ Justification: let C and D be two critical regions such that $C \subset D$.

Then $P_\theta (X \in C) \leq P_\theta (X \in D) \; \forall \; \theta \in \Omega$

For $\forall \; \theta \in \Omega_0$ $P_\theta (X \in C) \leq P_\theta (X \in D)$

For $\forall \; \theta \in \Omega_1$ $1-P_\theta (X \in C) \geq 1-P_\theta (X \in D)$

Hence by shrinking or enlarging a critical region we can decrease one type of error probability at the cost of increase in the error probability of another type.

Hence two type of error probabilities cannot be controlled simultaneously.]

**Testing Problem:**

Consider a population distribution with distribution function F(x ;θ) , $\theta \in \Omega$

Consider two hypotheses of interest H$_1$: $\theta \in \Omega_0$ and H$_2$: $\theta \in \Omega_1$ , where $\Omega_0 \cap \Omega_1 = \phi$

and $\Omega_0 \cup \Omega_1 = \Omega^* \subseteq \Omega$ .

here $\Omega^*$ is known as parameter space of interest. Let $(x_1, x_2, x_3, \ldots \ldots, x_n)$ be an observed sample

from the population distribution. The data provides information regarding population. Based

on the data and a testing rule to test or examine whether the data supports H$_1$ or H$_2$ and if the

data supports H$_1$ then we accept H$_1$

**The choice of Null Hypothesis:**

In the formulation of testing problem, the roles of H$_1$ and H$_2$ are not symmetric.

In testing of hypothesis, a statistician should be impartial and should have no brief for any

party; nor should he allow his personal view to influence the decision.

Consider an example, let $\mu_0$ and $\mu_1$ be average effects of drug manufacturers by an old process

and new process.

Then three hypotheses are possible: $\begin{cases} H_1 : \mu_0 = \mu_1 \\ H_2 : \mu_0 > \mu_1 \\ H_3 : \mu_0 < \mu_1 \end{cases}$

The last two hypotheses appeared to be biased as they reflect preferential attitude. The first

hypothesis suggests neutral or null attitude regarding the output. This neutral or Null attitude

of the statistician before the sample values are taken is the key to choose the null hypothesis.

Keeping in mind, the potential losses due to the wrong decision, the decision maker is

somewhat conservative in holding the null hypothesis as true, unless there is strong evidence

that it is false and to him consequences of wrongly rejecting the null hypothesis seems to be

more serious.

Hence, we denote by H$_0$: θ$\in \Omega_0$ , the null hypothesis; the hypothesis among H$_1$ and H$_2$ whose

false rejection is more serious.

Hence we are more concerned with the false rejection of true null hypothesis i.e Prob{Type-I-error}

**Why we take type -I-error probability in the simulation study to compare tests:**

In any testing problem the usual procedure of finding a test H$_0$ against H$_1$ is to restrict

that error probability which is more serious i.e the probability of type-I-error and then to

minimize the probability of type-II-error as far as possible.

Hence one selects a number $\alpha \in [0,1]$ and impose restriction on tests given by critical region W,

$P[X \in W|\theta] \leq \alpha$, for all $\theta \in \Omega_0$

Thus, the quantity $\alpha$ is called the level of significance of testing problem of testing $H_0$ vs $H_1$. The choice of level of significance of course depends on the experimenter himself. If he thinks that the false rejection of $H_0$ (type -I-error) is more serious error, then he will rather take a small value of α, say 0.01, 0.001 etc. On the other hand, if he thinks that this error is not so serious he will not mind taking a value of α high as α=0.05,0.1

So Now We will see the probability of type-I- errors in the above 4 tests for different choices of $n_1$ and $n_2$ and $\sigma_1^2, \sigma_2^2$

To obtain the results of classical t-test, Fiducial, Welch and Banerjee's tests, we use simulation consisting of 5000 runs for each of the sample size and parameter configurations. The Monte Carlo method is used for estimating the type I error rates.

The estimates of type I error rates of five tests are present in Table.

**Methodology:**

In order to find the empirical type -I- error probability we shall use a simulation technique as follows:

**STEP 1:** Consider the number of simulations be R, suppose R=1000; and let the sample size of the first sample be $n_1$. Then we will draw a sample of size R*$n_1$ at a time from the normal population with pre-specified parameter values. Then we shall store the sample in a matrix M (say) which has R columns and $n_1$ rows i.e each column of M will represent a particular Sample and we have R samples from the population in our hand.

**STEP 2:** Now we will compute the sample mean and sample variance for each sample, like We will get 1000 such sample means and sample variances.

**STEP 3:** Similarly, we shall perform the same work for the second normal population with Specified population parameters and sample size be $n_2$

**STEP 4:** Now we shall compute the Test statistic based on the sample observations under the

null hypothesis which will vary from sample to sample.

Let $T(X) = T(x_1,x_2,\ldots\ldots, x_n)$ be the observed value of the test statistic based on the sample $(x_1,x_2,\ldots\ldots, x_n)$ and $T_r(X)$ be the observed value of the test statistic in the $r^{th}$ simulation.

**STEP 5:** Now we shall compute the critical region for the 4 tests which will be different from each other for four tests described above.

**STEP 6:** For testing $H_0: \theta = \theta_0 \; vs \; H_1: \theta = \theta_1$, Probability of Type I error is defined as follows P{Type I error}=P{rejecting $H_0$ when $H_0$ is true}=P{$E_1$}

Empirical level can be obtained by calculating the proportion of times the event $E_1$ occurs i.e calculating the proportion of times, the true null hypothesis is rejected.

R-programming will be used to carry out all the necessary simulation procedures to obtain the Type I error probability.

Here our testing problem is:

$$H_0: \mu_1 = \mu_2 \; against \; H_1: \mu_1 \neq \mu_2$$

Where, $\mu_1$ is the mean of the first population and $\mu_2$ is the mean of the second population.

We have taken the level of significance (α) as 0.05 throughout our simulation study.

Type I error probabilities are represented in the tables below for different choices of $n_1$ and $n_2$.

In the first table we have taken the sample size of the populations be 5 and 5

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| TABLE 1: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05 | | | | |
| $(n_1, n_2) = (5, 5)$ | | | | |
| (5,5) | 0.0474 | 0.0226 | 0.0420 | 0.0228 |
| (5,10) | 0.0520 | 0.0246 | 0.0462 | 0.0246 |
| (5,25) | 0.0604 | 0.0318 | 0.0496 | 0.0318 |
| (5,40) | 0.0632 | 0.0356 | 0.0508 | 0.0358 |
| (5,60) | 0.0670 | 0.0390 | 0.0532 | 0.0390 |
| (5,90) | 0.0698 | 0.0428 | 0.0528 | 0.0424 |
| (5,120) | 0.0726 | 0.0438 | 0.0516 | 0.0434 |
| (5,150) | 0.0740 | 0.0438 | 0.0520 | 0.0438 |
| (5,200) | 0.0760 | 0.0454 | 0.0524 | 0.0450 |

In the second table we have taken the sample size of the populations be 5 and 10 respectively

| $(\sigma_1^2, \sigma_2^2)$ | **Fisher's t test** | **Fisher's fiducial approach** | **Welch's t-test** | **Banerjee's approach** |
|---|---|---|---|---|
| **TABLE 2: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05** | | | | |
| $(n_1, n_2) = (5, 10)$ | | | | |
| (5,5) | 0.0510 | 0.0358 | 0.0544 | 0.0344 |
| (5,10) | 0.0340 | 0.0324 | 0.0502 | 0.0318 |
| (5,25) | 0.0232 | 0.0334 | 0.0514 | 0.0332 |
| (5,40) | 0.0196 | 0.0390 | 0.0526 | 0.0386 |
| (5,60) | 0.0172 | 0.0410 | 0.0530 | 0.0408 |
| (5,90) | 0.0170 | 0.0450 | 0.0548 | 0.0448 |
| (5,120) | 0.0166 | 0.0456 | 0.0544 | 0.0456 |
| (5,150) | 0.0160 | 0.0454 | 0.0530 | 0.0454 |
| (5,200) | 0.0162 | 0.0478 | 0.0528 | 0.0478 |

In the next table we have taken the sample size of the populations be 10 and 5 respectively

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| **TABLE 3: Type 1 error probabilities for the Behrens-Fisher Problem when** $\mu_1 = \mu_2 = 20$ **and α=0.05** | | | | |
| $(n_1, n_2) = (10, 5)$ | | | | |
| (5,5) | 0.0442 | 0.0340 | 0.0490 | 0.0336 |
| (5,10) | 0.0772 | 0.0380 | 0.0526 | 0.0376 |
| (5,25) | 0.1244 | 0.0434 | 0.0540 | 0.0422 |
| (5,40) | 0.1504 | 0.0438 | 0.0534 | 0.0432 |
| (5,60) | 0.1650 | 0.0456 | 0.0536 | 0.0454 |
| (5,90) | 0.1810 | 0.0468 | 0.0538 | 0.0464 |
| (5,120) | 0.1888 | 0.0476 | 0.0534 | 0.0474 |
| (5,150) | 0.1938 | 0.0486 | 0.0534 | 0.0484 |
| (5,200) | 0.1998 | 0.0492 | 0.0524 | 0.0490 |

In the next table we have taken the sample size of the populations be 10 and 10 respectively

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| **TABLE 4: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05** | | | | |
| $(n_1, n_2) = (10, 10)$ | | | | |
| (5,5) | 0.0516 | 0.0408 | 0.0502 | 0.0400 |
| (5,10) | 0.0552 | 0.0424 | 0.0530 | 0.0418 |
| (5,25) | 0.0586 | 0.0464 | 0.0532 | 0.0454 |
| (5,40) | 0.0606 | 0.0474 | 0.0540 | 0.0458 |
| (5,60) | 0.0626 | 0.0484 | 0.0534 | 0.0474 |
| (5,90) | 0.0636 | 0.0490 | 0.0534 | 0.0486 |
| (5,120) | 0.0648 | 0.0504 | 0.0528 | 0.0496 |
| (5,150) | 0.0660 | 0.0500 | 0.0540 | 0.0498 |
| (5,200) | 0.0666 | 0.0510 | 0.0530 | 0.0510 |

Now we have taken $n_1$=10, $n_2$=25

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| **TABLE 5: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05** | | | | |
| $(n_1, n_2) = (10, 25)$ | | | | |
| (5,5) | 0.0504 | 0.0428 | 0.0496 | 0.0424 |
| (5,10) | 0.0276 | 0.0398 | 0.0482 | 0.0388 |
| (5,25) | 0.0134 | 0.0434 | 0.0494 | 0.0432 |
| (5,40) | 0.0096 | 0.0462 | 0.0500 | 0.0450 |
| (5,60) | 0.0074 | 0.0466 | 0.0504 | 0.0462 |
| (5,90) | 0.0070 | 0.0476 | 0.0516 | 0.0466 |
| (5,120) | 0.0064 | 0.0486 | 0.0514 | 0.0482 |
| (5,150) | 0.0060 | 0.0482 | 0.0502 | 0.0480 |
| (5,200) | 0.0058 | 0.0486 | 0.0510 | 0.0484 |

In the next table we have taken $n_1$=25, $n_2$=10

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|:---:|:---:|:---:|:---:|:---:|
| | | TABLE 6: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05 | | |
| | | $(n_1, n_2) = (25, 10)$ | | |
| (5,5) | 0.0550 | 0.0470 | 0.0558 | 0.0460 |
| (5,10) | 0.0952 | 0.0468 | 0.0536 | 0.0458 |
| (5,25) | 0.1516 | 0.0476 | 0.0510 | 0.0470 |
| (5,40) | 0.0810 | 0.0484 | 0.0512 | 0.0476 |
| (5,60) | 0.2000 | 0.0498 | 0.0506 | 0.0496 |
| (5,90) | 0.2130 | 0.0498 | 0.0508 | 0.0490 |
| (5,120) | 0.2208 | 0.0500 | 0.0508 | 0.0500 |
| (5,150) | 0.2248 | 0.0502 | 0.0506 | 0.0498 |
| (5,200) | 0.2298 | 0.0500 | 0.0502 | 0.0500 |

Next, we have taken $n_1$=25, $n_2$=25

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| TABLE 7: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05 | | | | |
| $(n_1, n_2) = (25, 25)$ | | | | |
| (5,5) | 0.0454 | 0.0410 | 0.0454 | 0.0402 |
| (5,10) | 0.0448 | 0.0404 | 0.0438 | 0.0394 |
| (5,25) | 0.0464 | 0.0422 | 0.0446 | 0.0420 |
| (5,40) | 0.0494 | 0.038 | 0.0462 | 0.0436 |
| (5,60) | 0.0502 | 0.0446 | 0.0466 | 0.0444 |
| (5,90) | 0.0514 | 0.0456 | 0.0466 | 0.0454 |
| (5,120) | 0.0528 | 0.0460 | 0.0468 | 0.0454 |
| (5,150) | 0.0524 | 0.0466 | 0.0472 | 0.0462 |
| (5,200) | 0.0534 | 0.0464 | 0.0472 | 0.0464 |

Next, we have taken $n_1$=25, $n_2$=50

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| **TABLE 8: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05** | | | | |
| $(n_1, n_2) = (25, 50)$ | | | | |
| (5,5) | 0.0482 | 0.0440 | 0.0470 | 0.0434 |
| (5,10) | 0.0278 | 0.0450 | 0.0490 | 0.0444 |
| (5,25) | 0.0156 | 0.0428 | 0.0468 | 0.0422 |
| (5,40) | 0.0116 | 0.0456 | 0.0468 | 0.0454 |
| (5,60) | 0.0100 | 0.0462 | 0.0492 | 0.0456 |
| (5,90) | 0.0084 | 0.0484 | 0.0496 | 0.0482 |
| (5,120) | 0.0082 | 0.0496 | 0.0512 | 0.0492 |
| (5,150) | 0.0080 | 0.0494 | 0.0498 | 0.0494 |
| (5,200) | 0.0080 | 0.0490 | 0.0494 | 0.0490 |

Next, we have taken n₁=50, n₂=25

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| **TABLE 9: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05** | | | | |
| $(n_1, n_2) = (50, 25)$ | | | | |
| (5,5) | 0.0512 | 0.0506 | 0.0542 | 0.0496 |
| (5,10) | 0.0818 | 0.0498 | 0.0520 | 0.0494 |
| (5,25) | 0.1238 | 0.0504 | 0.0514 | 0.0500 |
| (5,40) | 0.1392 | 0.0512 | 0.0520 | 0.0508 |
| (5,60) | 0.1470 | 0.0514 | 0.0520 | 0.0504 |
| (5,90) | 0.1548 | 0.0508 | 0.0510 | 0.0508 |
| (5,120) | 0.1588 | 0.0514 | 0.0520 | 0.0512 |
| (5,150) | 0.1614 | 0.0512 | 0.0516 | 0.0512 |
| (5,200) | 0.1650 | 0.0502 | 0.0504 | 0.0502 |

Next, we have taken $n_1$=50, $n_2$=50

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| **TABLE 10: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05** | | | | |
| $(n_1, n_2) = (50, 50)$ | | | | |
| (5,5) | 0.0462 | 0.0440 | 0.0462 | 0.0436 |
| (5,10) | 0.0482 | 0.0464 | 0.0480 | 0.0462 |
| (5,25) | 0.0484 | 0.0460 | 0.0474 | 0.0460 |
| (5,40) | 0.0492 | 0.0468 | 0.0480 | 0.0464 |
| (5,60) | 0.0474 | 0.0452 | 0.0454 | 0.0452 |
| (5,90) | 0.0476 | 0.0462 | 0.0466 | 0.0460 |
| (5,120) | 0.0490 | 0.0472 | 0.0478 | 0.0472 |
| (5,150) | 0.0488 | 0.0474 | 0.0474 | 0.0474 |
| (5,200) | 0.0492 | 0.0468 | 0.0470 | 0.0468 |

In the next table $n_1$=50, $n_2$=100

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| **TABLE 11: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05** | | | | |
| $(n_1, n_2) = (50, 100)$ | | | | |
| (5,5) | 0.0466 | 0.0432 | 0.0452 | 0.0430 |
| (5,10) | 0.0244 | 0.0448 | 0.0464 | 0.0442 |
| (5,25) | 0.0128 | 0.0442 | 0.0450 | 0.0440 |
| (5,40) | 0.0100 | 0.0438 | 0.0448 | 0.0438 |
| (5,60) | 0.0086 | 0.0434 | 0.0438 | 0.0434 |
| (5,90) | 0.0072 | 0.0432 | 0.0438 | 0.0432 |
| (5,120) | 0.0068 | 0.0432 | 0.0436 | 0.0430 |
| (5,150) | 0.0064 | 0.0424 | 0.0432 | 0.0424 |
| (5,200) | 0.0054 | 0.0424 | 0.0430 | 0.0422 |

Next, we have $n_1$=100, $n_2$=50

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| TABLE 12: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05 | | | | |
| $(n_1, n_2) = (100, 50)$ | | | | |
| (5,5) | 0.0388 | 0.0364 | 0.0382 | 0.0364 |
| (5,10) | 0.0710 | 0.0412 | 0.0432 | 0.0408 |
| (5,25) | 0.1164 | 0.0442 | 0.0450 | 0.0442 |
| (5,40) | 0.1312 | 0.0452 | 0.0458 | 0.0452 |
| (5,60) | 0.1410 | 0.0466 | 0.0466 | 0.0462 |
| (5,90) | 0.1490 | 0.0460 | 0.0464 | 0.0460 |
| (5,120) | 0.1538 | 0.0470 | 0.0476 | 0.0470 |
| (5,150) | 0.1566 | 0.0484 | 0.0486 | 0.0484 |
| (5,200) | 0.1598 | 0.0488 | 0.0490 | 0.0488 |

Lastly, we have taken $n_1$=100, $n_2$=100

| $(\sigma_1^2, \sigma_2^2)$ | Fisher's t test | Fisher's fiducial approach | Welch's t-test | Banerjee's approach |
|---|---|---|---|---|
| **TABLE 13: Type 1 error probabilities for the Behrens-Fisher Problem when $\mu_1 = \mu_2 = 20$ and α=0.05** | | | | |
| $(n_1, n_2) = (100, 100)$ | | | | |
| (5,5) | 0.0430 | 0.0428 | 0.0430 | 0.0426 |
| (5,10) | 0.0462 | 0.0456 | 0.0462 | 0.0456 |
| (5,25) | 0.0484 | 0.0472 | 0.0480 | 0.0470 |
| (5,40) | 0.0500 | 0.0494 | 0.0496 | 0.0492 |
| (5,60) | 0.0486 | 0.0478 | 0.0480 | 0.0478 |
| (5,90) | 0.0482 | 0.0474 | 0.0474 | 0.0474 |
| (5,120) | 0.0484 | 0.0474 | 0.0478 | 0.0474 |
| (5,150) | 0.0484 | 0.0470 | 0.0470 | 0.0470 |
| (5,200) | 0.0486 | 0.0472 | 0.0472 | 0.0472 |

## Conclusion and findings:

The estimates of type I error rates of four tests are presented above. We have the following numerical results.

- ❖ As long as the variances were homogeneous, the classical Fisher's t test seems to have type I error rate quite close to α, the nominal level and type I error rates of the remaining three tests are less than α.

- ❖ When departure from equality of variances increases this type-I-error probability poorly deviates from the nominal level which depends on the sample sizes of the two populations.

- ❖ **Case 1: $n_1 < n_2$**

   When the sample size of the first population is less than the sample size of the second population and as the deviation of $\sigma_2^2$ from $\sigma_1^2$ increases i.e the heteroscedasticity increases, then the Type I error probability for the classical Fisher's t test will decrease rapidly and deviates from α, Fiducial approach by Fisher and Banerjee's approach performs almost same. Type I error of Welch t test increases very slowly.

- ❖ **Case 2: $n_1 > n_2$**

   When $n_1 > n_2$ then for the Fisher t test the probability of type I error increases rapidly and poorly deviates from α as the heteroscedasticity increases. Probability of type I error is almost equal for Fisher's Fiducial approach and Banerjee's approach and decreases in Welch test as the difference in variance increases.

- ❖ **Case 3: $n_1 = n_2$**

   When $n_1 = n_2$, type I error rate for Fisher t test is quite close to α and here also probabilities are same for Fisher's Fiducial approach and Banerjee's approach and Welch t test and those are close to α, the nominal value, as the difference in the two variances increases.

- ❖ The Welch t test (WS) has satisfactory type I error rate regardless of sample sizes and unequal variances. Fisher's fiducial approach and Banerjee's approach performs almost same in all case irrespective of sample size and unequal variances.

- ❖ Fisher's fiducial approach and Banerjee's test seems to be very conservative, when sample sizes are small. The fiducial approach, Welch test and Banerjee's test have similar results when the sample sizes are large.

- ❖ For unequal sample sizes, the Fisher t test is far worse than type I error rates of the Fiducial, Welch and Banerjee's tests. But its type I error rate close when sample sizes are equal and large.

- ❖ This also show that Fisher's t test is not robust against the departure from homoscedasticity.

Example: When $n_1$=50 and $n_2$=25 $(n_1 > n_2)$, and $\mu_1 = \mu_2 = 20$ then for different choices of variances the type I error rate of fisher's t test poorly deviates from the nominal level α (0.05). From Table 9 we have the following probability of type I errors of Fisher's t test for different pairs of variances

| $(\sigma_1^2, \sigma_2^2)$ | (5,5) | (5,10) | (5,25) | (5,40) | (5,60) | (5,90) | (5,120) | (5,150) | (5,200) |
|---|---|---|---|---|---|---|---|---|---|
| **Prob of Type-I-error** | 0.0512 | 0.0818 | 0.1238 | 0.1392 | 0.1470 | 0.1548 | 0.1588 | 0.1614 | 0.1650 |

And when $n_1 < n_2$ ; Say, $n_1$=25, $n_2$=50 and $\mu_1 = \mu_2 = 20$ then the fisher's t test becomes more conservative in rejecting H0

From Table 8 We have the following probability of type I errors of Fisher's t test for the above pair of variances

| $(\sigma_1^2, \sigma_2^2)$ | (5,5) | (5,10) | (5,25) | (5,40) | (5,60) | (5,90) | (5,120) | (5,150) | (5,200) |
|---|---|---|---|---|---|---|---|---|---|
| **Prob of Type-I-error** | 0.0482 | 0.0278 | 0.0156 | 0.0116 | 0.0100 | 0.0084 | 0.0082 | 0.0080 | 0.0080 |

## Comparison of above tests by checking power through Power Curve:

Now we will compute the power of the above 4 tests and represent it graphically through Power Curve

Power is the probability of the rejection of the false null hypothesis and it is an important criterion for good test. Generally, a test will be considered as a good test when the power is as much as possible. Clearly the power of the test is the probability of taking one of the two correct decisions arising in the testing problem. Now we will define Power Function.

## Power Function:

Let $\gamma$ be a test of the null hypothesis H0. The power function of the test $\gamma$, denoted by $\pi_\gamma(\theta)$, is defined to be the probability that H0 is rejected when the distribution from which the sample was obtained was parameterized by θ.

The power function will play the same role in hypothesis testing that mean-squared error played in estimation. It will usually be our standard in assessing the goodness of a test or in

comparing two competing tests. An ideal power function, of course, is a function that is 0 for those θ corresponding to the null hypothesis and is unity for those θ corresponding to the alternative hypothesis. The idea is that you do not want to reject $H_0$ if $H_0$ is true and you do want to reject $H_0$ when $H_0$ is false.

**Importance of Power Function in hypothesis testing:**

Considering as a function of $\theta \in \Omega$ (the parameter space) , the probability $P[X \in W|\theta] = \beta_\theta$, is called the power function of the test.

$$\text{For } \theta \in \Omega_0 , \ \beta_\theta = P[X \in W|\theta] = \text{Probability of Type I error}$$

$$\text{For } \theta \in \Omega_1 , \ \beta_\theta = P[X \in W|\theta] = 1 - P[X \in W^c|\theta] = 1 - \text{Probability of Type II error}$$

So Now We will compute the empirical power of the above 4 tests for different choices of $n_1$ and $n_2$ and $\sigma_1^2, \sigma_2^2$

To obtain the results of classical t-test, Fiducial, Welch and Banerjee's tests, we use simulation consisting of 1000 runs for each of the sample size and parameter configurations. The Monte Carlo method is used for estimating the power.

**Methodology:**

In order to find the empirical power, we shall use a simulation technique as follows:

**STEP 1:** Consider the number of simulations be R, suppose R=1000; and let the sample size of the first sample be $n_1$. Then we will draw a sample of size $R*n_1$ at a time from the normal population with pre-specified parameter values. Then we shall store the sample in a matrix M (say) which has R columns and $n_1$ rows i.e each column of M will represent a particular Sample and we have R samples from the population in our hand.

**STEP 2:** Now we will compute the sample mean and sample variance for each sample, like We will get 1000 such sample means and sample variances.

**STEP 3:** Similarly, we shall perform the same work for the second normal population with Specified population parameters and sample size be $n_2$

**STEP 4:** Now we shall compute the Test statistic based on the sample observations under the null hypothesis which will vary from sample to sample.

Let T(X) = T($x_1, x_2, \ldots\ldots, x_n$) be the observed value of the test statistic based on the sample ($x_1, x_2, \ldots\ldots, x_n$) and $T_r(X)$ be the observed value of the test statistic in the $r^{th}$ simulation.

**STEP 5:** Now we shall compute the critical region for the 4 tests which will be different

from each other for four tests described above.

**STEP 6:** For testing $H_0: \theta = \theta_0 \ vs \ H_1: \theta = \theta_1$, power of a test is defined as follows

Power of the test =Prob {rejecting $H_0$ when $H_0$ is false} =1-Prob {accepting $H_0$ when $H_0$ is false} =1-P{$E_2$}

Empirical power can be obtained by calculating the proportion of times the event $E_2$ occurs

or calculating the proportion of times, the false null hypothesis is rejected.

R-programming will be used to carry out all the necessary simulation procedures to obtain the

Empirical power.

Here our testing problem is:

$$H_0: \mu_1 = \mu_2 \ against \ H_1: \mu_1 \neq \mu_2$$

Where $\mu_1$ is the mean of the first population and $\mu_2$ is the mean of the second population.

We have taken the level of significance (α) as 0.05 throughout our simulation study.

### An important observation comparing the Power of the tests

The problem arises when we are comparing the fisher's t test with Fiducial, Welch and

Banerjee's test. Since the empirical levels of the Fisher's t test poorly deviates from the

nominal level (0.05),

When $n_1 > n_2$, the test become more liberal in favour of rejecting $H_0$ likewise when $n_1 < n_2$,

The test becomes more conservative in rejecting $H_0$.

**Example:** When $n_1$=25 and $n_2$=10 $(n_1 > n_2)$, and $\mu_1 = \mu_2 = 20$ then for different choices of

variances the type I error rate of fisher's t test poorly deviates from the nominal level α (0.05).

From Table 6 we have the following probability of type I errors of Fisher's t test for different

pairs of variances

| $(\sigma_1^2, \sigma_2^2)$ | (5,5) | (5,10) | (5,25) | (5,40) | (5,60) | (5,90) | (5,120) | (5,150) | (5,200) |
|---|---|---|---|---|---|---|---|---|---|
| **Prob of Type-I-error** | 0.0550 | 0.0952 | 0.1516 | 0.0810 | 0.2000 | 0.2130 | 0.2208 | 0.2248 | 0.2298 |

Likewise, when $n_1$=10 and $n_2$=25 $(n_1 < n_2)$, and $\mu_1 = \mu_2 = 20$ then for different choices of

variances the type I error rate of Fisher's t test is vey much less than the nominal level

α (0.05)

From Table 5 we have the following probability of type I errors of Fisher's t test for different pairs of variances

| $(\sigma_1^2, \sigma_2^2)$ | (5,5) | (5,10) | (5,25) | (5,40) | (5,60) | (5,90) | (5,120) | (5,150) | (5,200) |
|---|---|---|---|---|---|---|---|---|---|
| **Prob of Type-I-error** | 0.0504 | 0.0276 | 0.0134 | 0.0096 | 0.0074 | 0.0070 | 0.0064 | 0.0060 | 0.0058 |

**Remark:** For a testing problem $H_0: \theta \in \Omega_0 \ vs \ H_1: \theta \in \Omega_1$ , with a chosen level of significance α, there are several tests and a test W may have size > α, then the test W will not be considered for construction of a good test. So, this shows that Fisher's t test is not no more robust against the departure from homoscedasticity.

Since the empirical levels of the Fisher's t test poorly fluctuate from the nominal level, it is not advisable to use it for power comparison with the other tests described above. Thus We discard Fisher's t test and work with the other tests proposed above mainly

**Graphs:** We obtain the power curves for Fiducial, Welch and Banerjee's test for different pairs of $(n_1, n_2)$ as the Welch's and Banerjee's tests are not permutation invariant i.e depends upon the sample sizes. Here the testing problem is $H_0: \mu_1 = \mu_2 = 0$ against $H_1: \mu_1 \neq \mu_2$
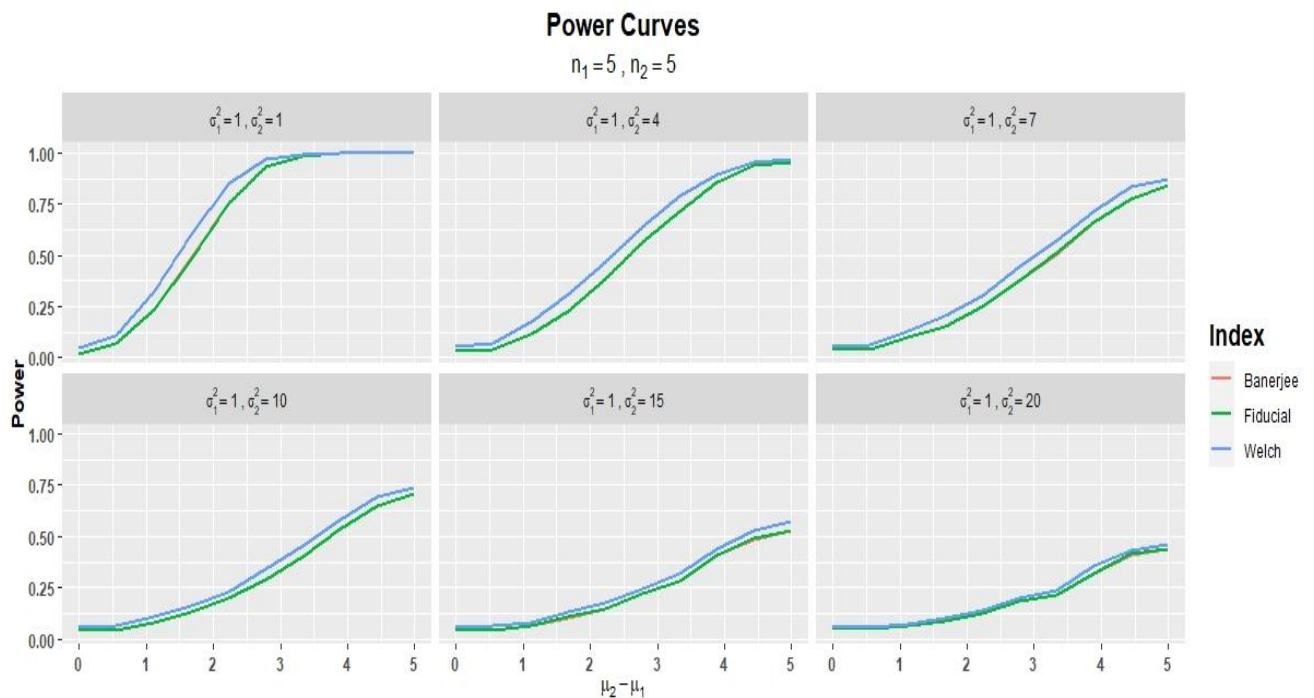


*Figure 1: $n_1 = 5, n_2 = 5$*
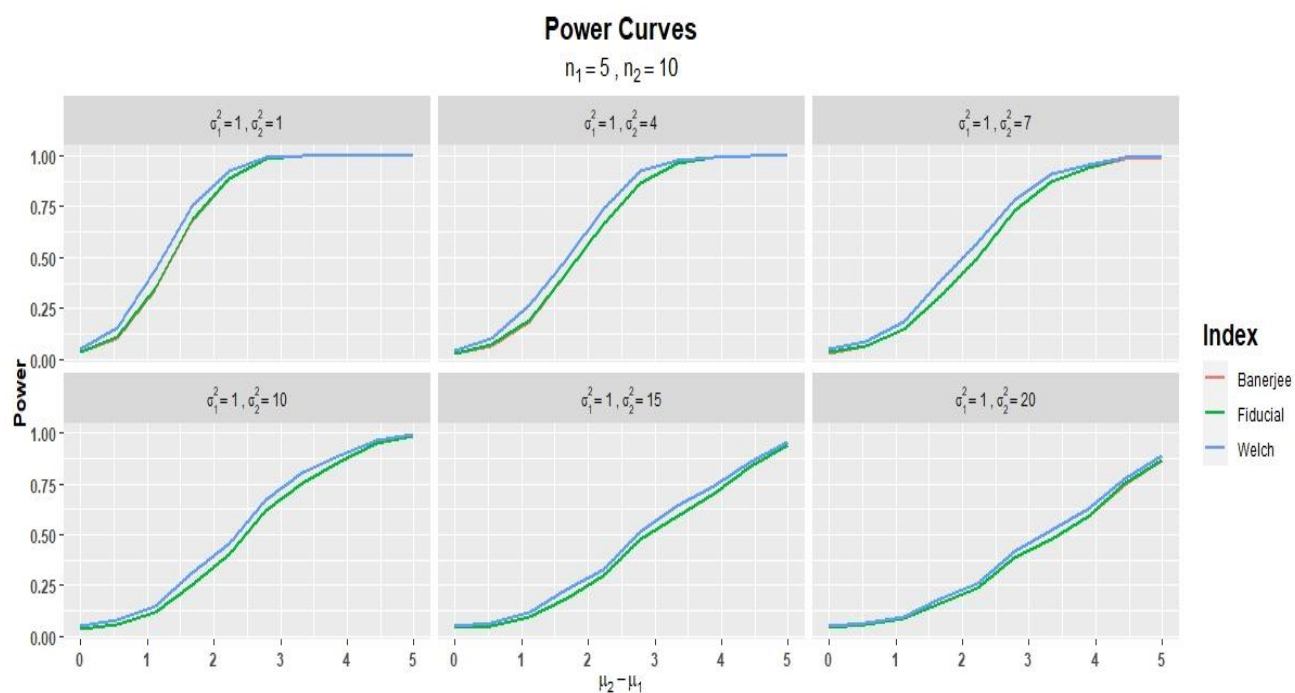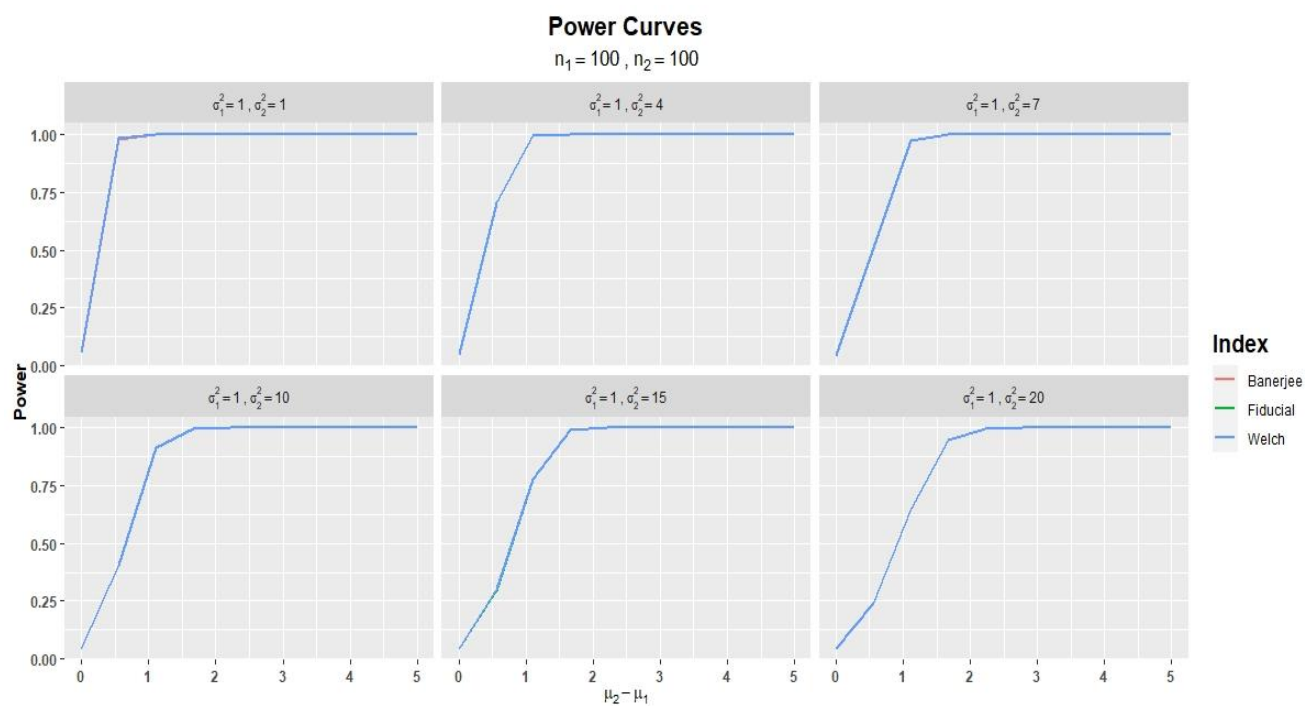
**Figure 2:** $n_1 = 5, n_2 = 10$
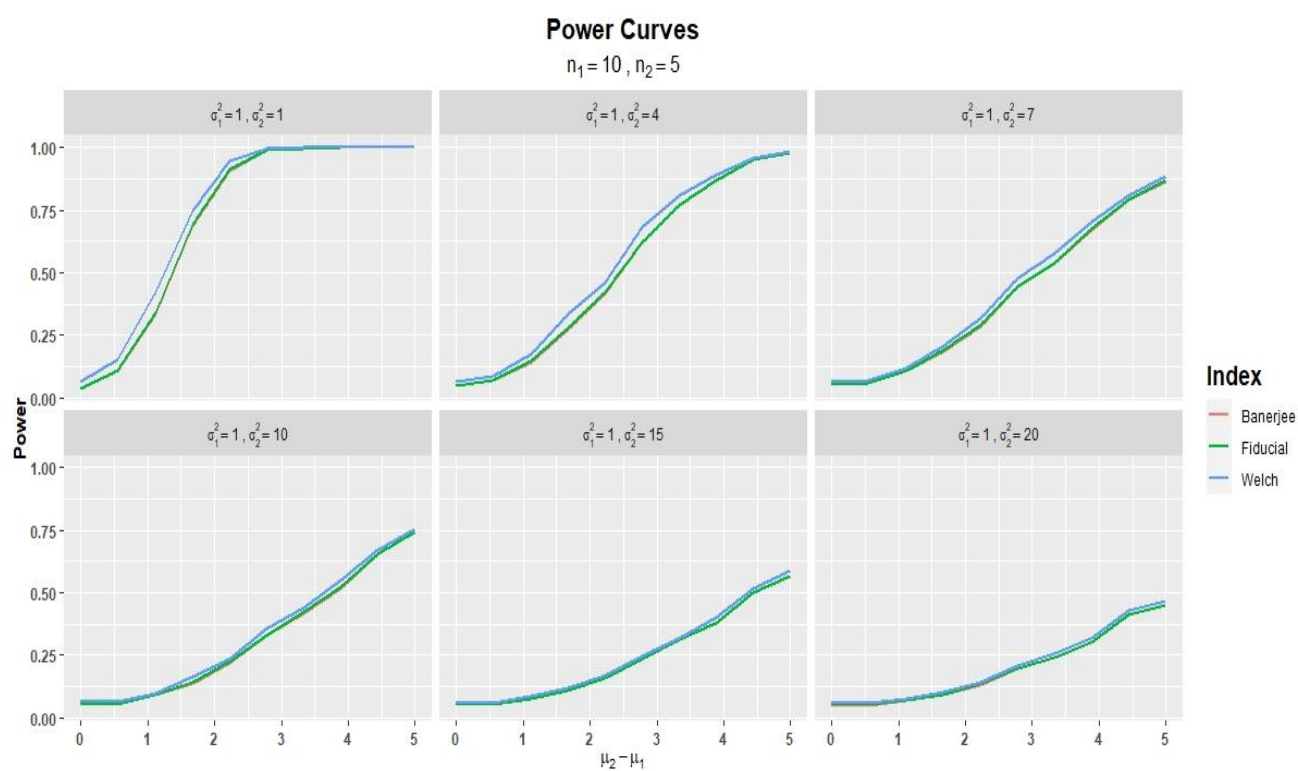


**Figure 3:** $n_1 = 10, n_2 = 5$
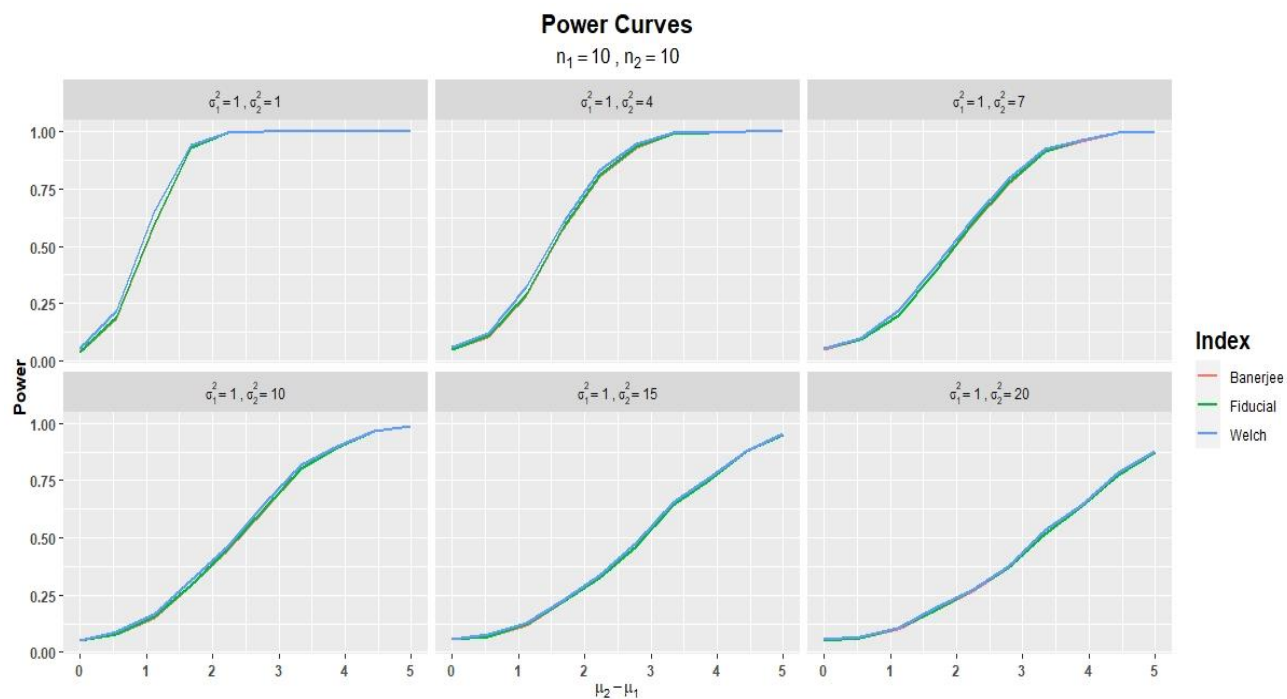
*Figure 4:* $n_1 = 10, n_2 = 10$
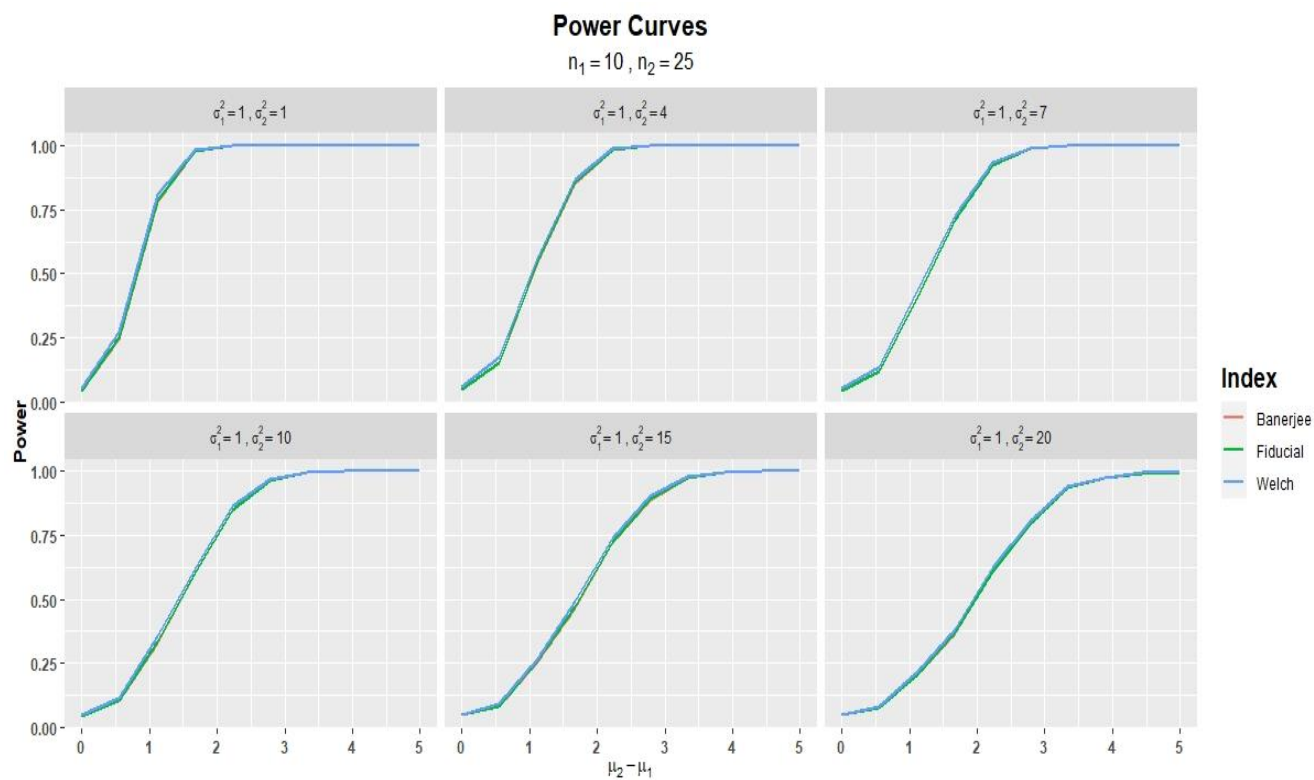


*Figure 5:* $n_1 = 10, n_2 = 25$
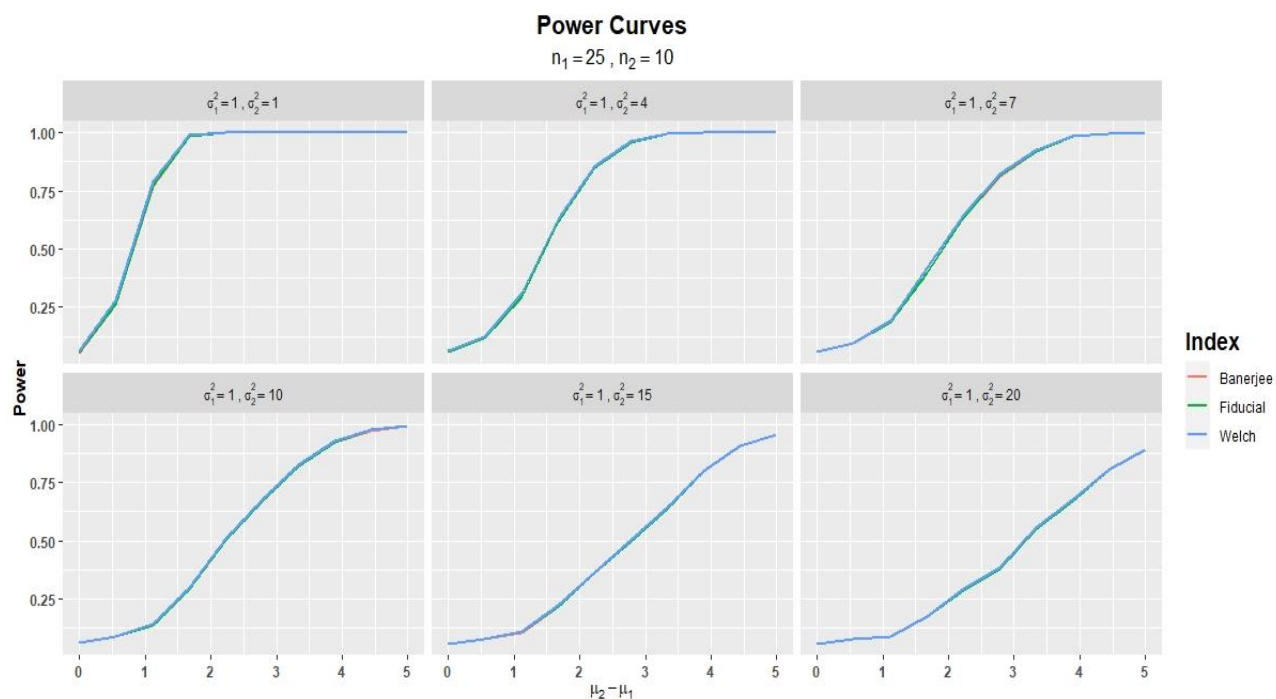
*Figure 6: $n_1 = 25, n_2 = 10$*



*Figure 7: $n_1 = 25, n_2 = 25$*

**Power Curves**

$n_1 = 25$ , $n_2 = 50$



*Figure 8: $n_1 = 25, n_2 = 50$*

**Power Curves**
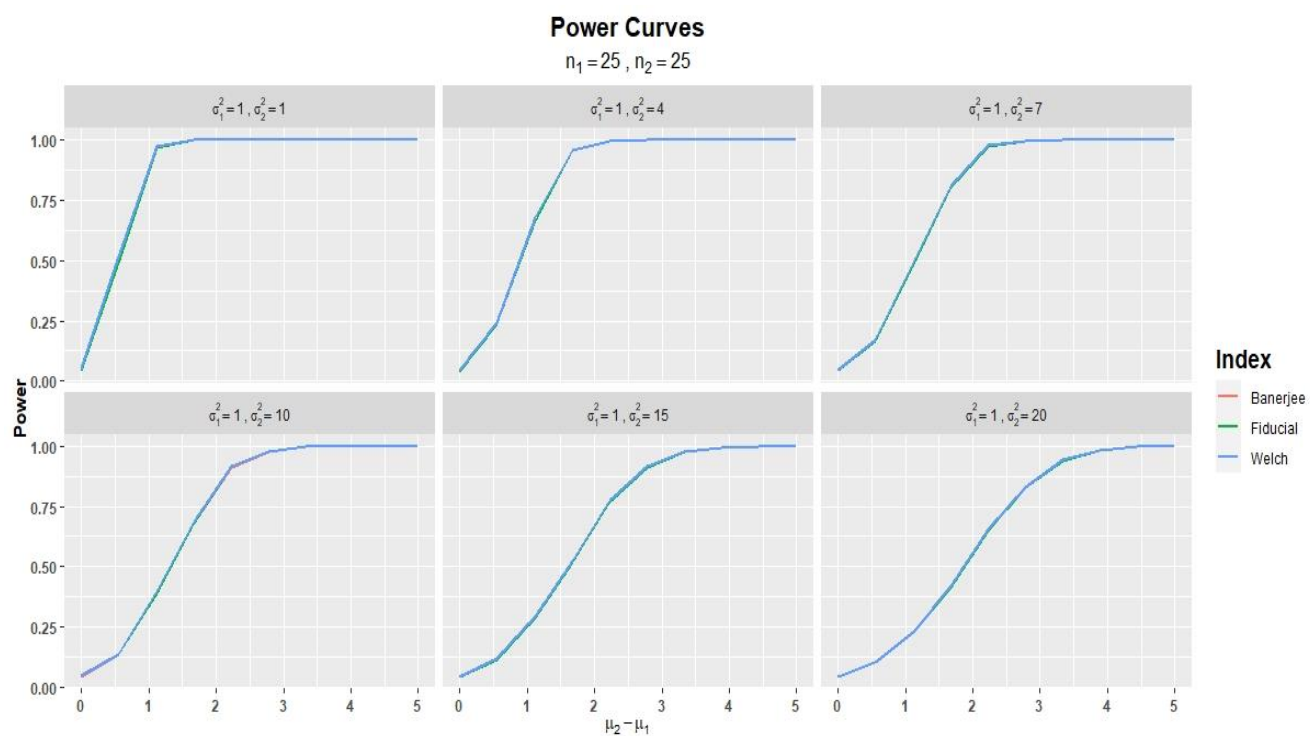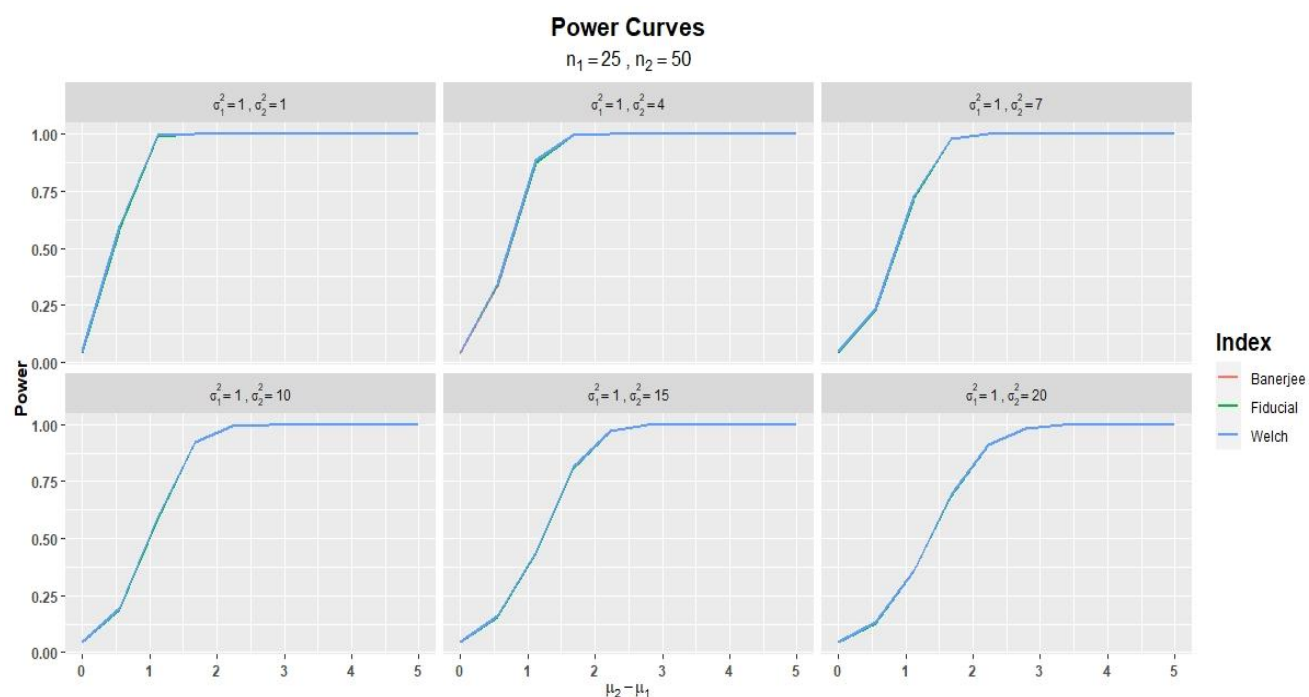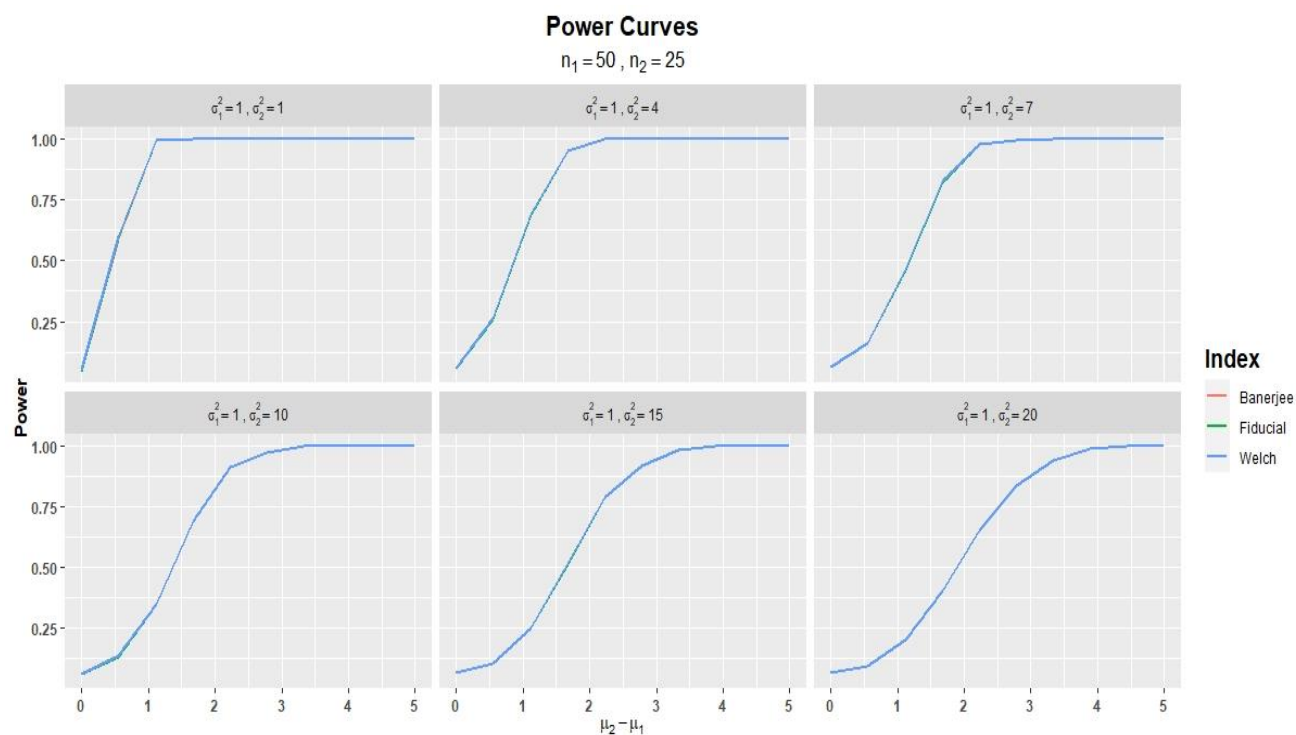
$n_1 = 50$ , $n_2 = 25$



*Figure 9: $n_1 = 50, n_2 = 25$*
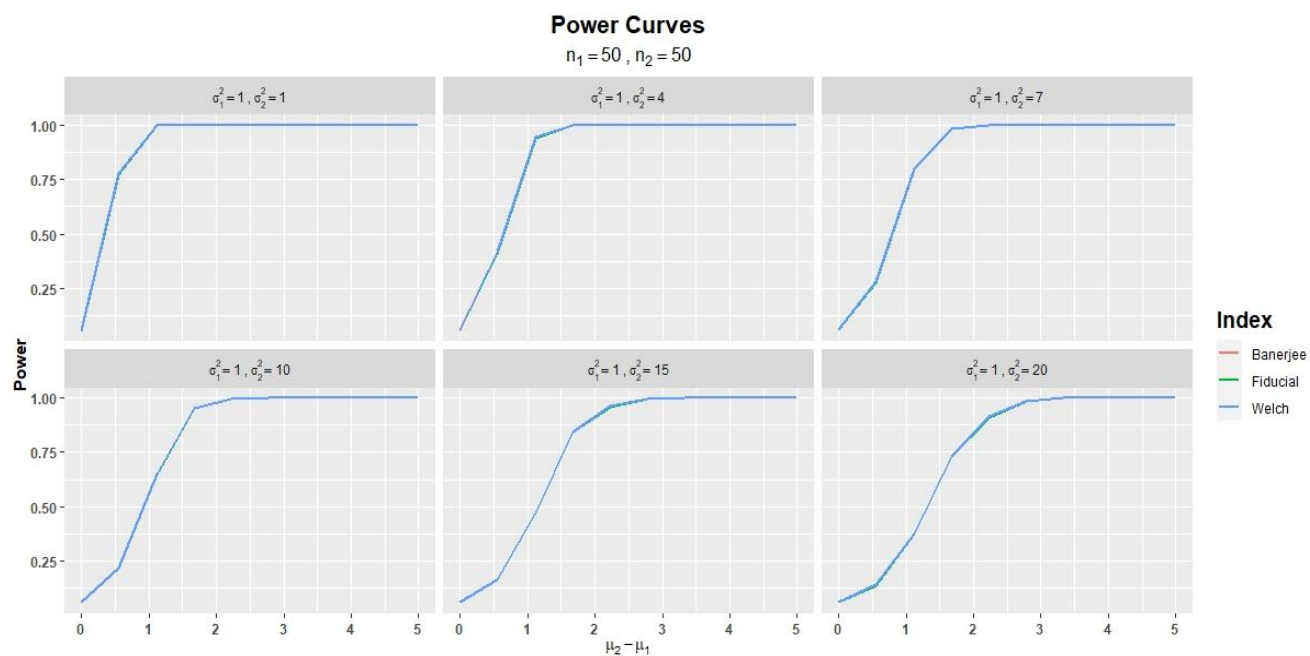
*Figure 10:* $n_1 = 50, n_2 = 50$
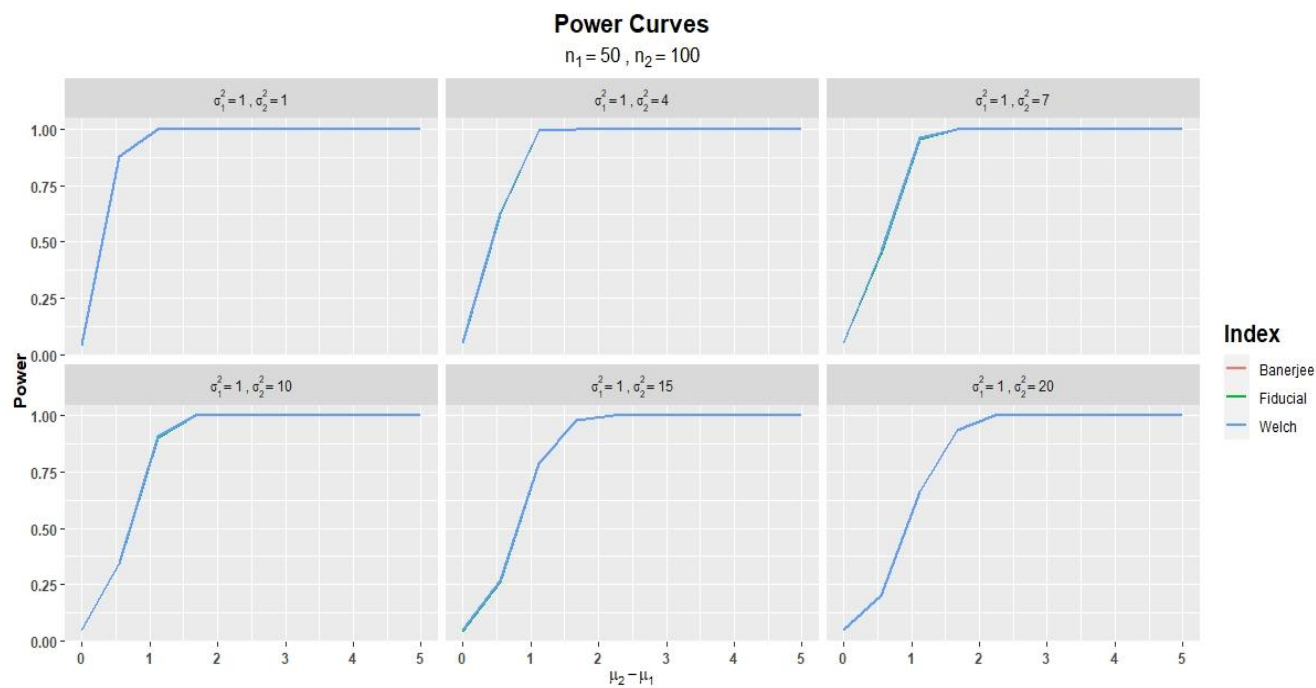


*Figure 11:* $n_1 = 50, n_2 = 100$
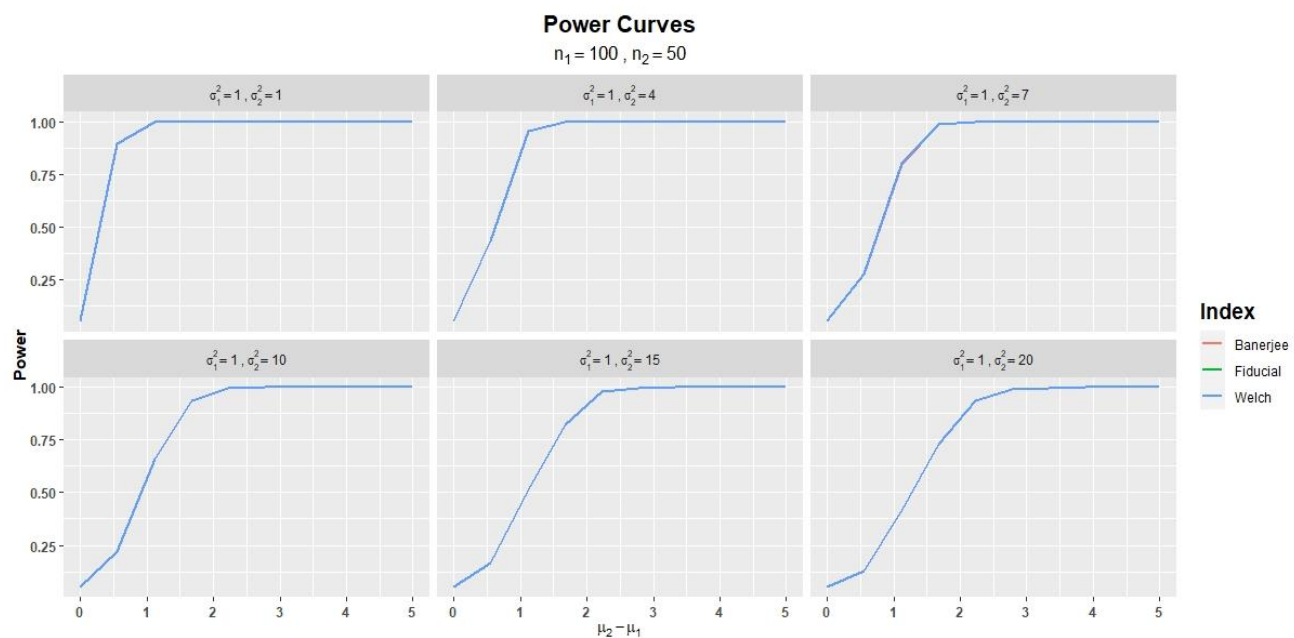
*Figure 12:* $n_1 = 100, n_2 = 50$



*Figure 13:* $n_1 = 100, n_2 = 100$

## **Conclusion and findings:**

❖ It is noticeable that the three tests namely Fiducial test, Welch test and Banerjee's test are not permutation invariant i.e after interchanging the role of $n_1$ and $n_2$ the performance of the tests will change i.e these tests are not symmetric about $n_1$ and $n_2$ and this is a major limitation of these three tests.

❖ The power of Banerjee's test and Fiducial test is more or less same in all the cases regardless of sample sizes and unequal variances.

❖ For small sample sizes, as the departure from the homoscedasticity increases the power of all three tests is very low.

❖ Welch's approach seems to be more powerful than fiducial approach and Banerjee's approach for small sample sizes.

❖ For large sample sizes three tests perform almost same for homogeneity as well as heterogeneity in variances.

❖ Fiducial approach and Banerjee's approach seems to be more conservative than Welch in rejecting the false null hypothesis.

❖ For moderate sample sizes as the heteroscedasticity increases the slope of the power curve decreases for all three tests and robustness of the tests decrease.

## **Limitations:**

One of the major limitations of these tests that these tests are not permutation invariant i.e due to interchanging the role of $n_1$ and $n_2$, the performance of the tests vary. The difficulty with the Behrens-Fisher problem is that exact solutions are not available satisfactorily because nuisance parameters are present in this problem. 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. In this review, we have focused on some of the parametric solutions to the Behrens-Fisher problem. Nonparametric or distribution-free solutions have also been proposed for this type of problem.

## References

- https://www.researchgate.net/publication/313652349_A_Simulation_Study_on_Tests_for_the_Behrens-Fisher_Problem

- https://files.eric.ed.gov/fulltext/ED393866.pdf

- http://library.isical.ac.in:8080/jspui/bitstream/10263/265/1/60.E05.pdf

- https://www.researchgate.net/publication/301292970_Welch's_t_test

- www.wikipedia.org

## Acknowledgement