Commun.Fac.Sci.Univ.Ank.Series A1
Volume 59, Number 2, Pages 15–34 (2010)
ISSN 1303–5991

A SIMULATION STUDY ON TESTS FOR ONE-WAY ANOVA UNDER THE UNEQUAL VARIANCE ASSUMPTION

ESRA YIĞIT AND FIKRI GÖKPINAR

ABSTRACT. The classical F-test to compare several population means depends on the assumption of homogeneity of variance of the population and the normality. When these assumptions especially the equality of variance is dropped, the classical F-test fails to reject the null hypothesis even if the data actually provide strong evidence for it. This can be considered a serious problem in some applications, especially when the sample size is not large. To deal with this problem, a number of tests are available in the literature. In this study, the Brown-Forsythe, Weerahandi's Generalized F, Parametric Bootstrap, Scott-Smith, One-Stage, One-Stage Range, Welch and Xu-Wang's Generalized F-tests are introduced and a simulation study is performed to compare these tests according to type-1 errors and powers in different combinations of parameters and various sample sizes.

1. INTRODUCTION

In applied statistics an experimenter wants to compare two or more populations measured using independent samples. The classical F- (CF) test is used under the assumption that the populations have normal distributions with the same variances. In this paper we consider the problem of comparing the means of k populations with the assumption of heteroscedastic variances.

The CF test fails to reject the null hypothesis even for large samples when the population variances are unequal. This is a serious problem, especially for biomedical experiments in which one does not usually have large samples. In such applications each data point can be so vital and expensive. Alternative methods are developed due to this problem. Some of these test statistics' distribution is not known and the p-value can be found by simulation (Weerahandi, 1995; Weerahandi, 2004). There

Received by the editors June 25, 2010, Accepted: October. 26, 2010.
2000 *Mathematics Subject Classification.* Primary 05C38, 15A15; Secondary 05A15, 15A18.
Key words and phrases. Brown-Forsythe test, Generalized F-test, Parametric Bootstrap test, Scott- Smith test, One-Stage test, One-Stage Range Test, Classic F-test, Welch test, Xu-Wang test.

are a large number of approximate tests (Chen and Chen, 1998; Chen, 2001; Tsui and Weerahandi, 1989; Krishnamoorthy et al., 2006; Xu and Wang, 2007a, 2007b) and exact tests (Bishop and Dudewicz, 1981; Welch, 1951; Scott and Smith, 1971; Brown and Forsythe, 1974) in the literature. In practice, some exact procedures such as the CF, Welch (W), Scott-Smith (SS) and Brown-Forsythe (BF) tests are widely used. Alternative tests have been applied to solve a number of problems when conventional methods are difficult to apply or fail to provide exact solutions.

In this paper we carry out a simulation study to compare the size performance of the CF, W, SS, BF, Chen-Chen's One Stage (OS), Chen-Chen's One Stage Range (OSR), Weerahandi's Generalized F (GF), Xu-Wang's Generalized F (XW) and Parametric Bootstrap (PB) tests when population variances are unequal in one-way ANOVA problems. The type-I error rates and powers of the tests are compared using Monte Carlo simulation using various sample sizes and under various parameter combinations.

2. TESTS FOR ONE-WAY ANOVA

Let $X_{i1},\ldots,X_{in_i}$ be a random sample from $N(\mu_i,\, \sigma_i^2)$ $i=1,\ldots,k$. The problem of interest involves testing

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad H_1 : \text{ Not all } \mu_i\text{s are equal} \quad i = 1,\ldots,k \tag{2.1}$$

The standardized between-group sum of squares and the standardized error sum of squares are given in (2.2) and (2.3) when σ_i^2s are unequal.

$$\tilde{S}_b = \tilde{S}_b\left(\sigma_1^2,\ldots,\sigma_k^2\right) = \sum_{i=1}^{k} \frac{n_i \bar{X}_i^2}{\sigma_i^2} - \frac{\left(\sum_{i=1}^{k} \frac{n_i \bar{X}_i}{\sigma_i^2}\right)^2}{\sum_{i=1}^{k} \frac{n_i}{\sigma_i^2}} \tag{2.2}$$

$$\tilde{S}_e = \sum_{i=1}^{k} \frac{n_i S_i^2}{\sigma_i^2} \tag{2.3}$$

Most of the test statistics to test the equality of means under heteroscedasticity are based on the standardized between-group sum of squares and standardized error sum of squares. In the rest of this section test statistics are briefly introduced. In this section the W, SS and BF tests, whose distribution can be obtained theorically, are given. GF test and the XW test based on the generalized F-test, whose p-values are obtained by simulation, are given. The OS and OSR tests developed by Chen and Chen (1998, 2001) based on Bishop and Dudewicz's (1981) two-stage procedure are investigated. Finally, the PB test developed by Krishnamoorthy et al. (2006) is discussed.

The Welch Test

If $w_i = \frac{n_i}{S_i^2}$, Welch (1951) gives the following test statistics.

$$W = \frac{\tilde{S}_b\left(S_1^2, \ldots, S_k^2\right) \Big/ (k-1)}{1 + \frac{2(k-2)}{k^2-1} \sum_{i=1}^{k} \frac{1}{n_i - 1} \left(1 - \frac{w_i}{\sum w_j}\right)^2} = \frac{\sum_{i=1}^{k} w_i \left[\left(\bar{X}_i - \bar{X}\right)^2 \Big/ (k-1)\right]}{1 + \frac{2(k-2)}{k^2-1} \sum_{i=1}^{k} \frac{1}{n_i - 1} \left(1 - \frac{w_i}{\sum w_j}\right)^2} \tag{2.4}$$

If H_0 is true, then the distribution of W is $F_{k-1,f}$ where

$$f = \frac{1}{\frac{3}{k^2-1} \sum_{i=1}^{k} \frac{1}{n_i - 1} \left(1 - \frac{w_i}{\sum w_j}\right)^2}$$

For a given level α, and an observed value W_h of W, this test rejects the H_0 in (2.1) whenever the p-value is given as $P\left(F_{k-1,f} \rangle W_h\right) \langle \alpha$.

The Scott-Smith Test

If $S_i^{*2} = \frac{n_i - 1}{n_i - 3} S_i^2$, Scott and Smith (1971) give the following test statistics.

$$F_s = \sum_{i=1}^{k} \frac{n_i \left(\bar{X}_i - \bar{X}\right)^2}{S_i^{*2}}$$

If H_o is true, then the distribution of F_s is χ_k^2. For a given level α, and an observed value fs of F_s, this test rejects the H_0 in (2.1) whenever the p-value is given as $P\left(F_s \rangle f_s\right) \langle \alpha$.

The Brown-Forsythe Test

Brown and Forsythe (1974) give the following test statistics.

$$B = \sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{X}\right)^2 \Bigg/ \sum_{i=1}^{k} \left(1 - \frac{n_i}{n}\right) S_i^2$$

If H_0 is true, then the distribution of B is $F_{k-1,v}$; where

$$v = \left[\sum_{i=1}^{k} \left(1 - \frac{n_i}{n}\right) S_i^2\right]^2 \Bigg/ \sum_{i=1}^{k} \frac{\left(1 - \frac{n_i}{n}\right)^2 S_i^4}{n_i - 1}$$

For a given level α, and an observed value B_h of B, this test rejects the H_0 in (2.1) whenever the p-value is given as $P\left(F_{k-1,v} \rangle B_h\right) \langle \alpha$.

Weerahandi's Generalized F-test

The sample variances (MLEs) of the k populations are denoted by S_i^2, where

$$S_i^2 = \frac{1}{n_i}\sum_{i=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2.$$

Define

$$B_j = \frac{\left(\sum_{i=1}^{j}\frac{n_i S_i^2}{\sigma_i^2}\right)}{\left(\sum_{i=1}^{j+1}\frac{n_i S_i^2}{\sigma_i^2}\right)}, \qquad j = 1, \ldots, k-1$$

It follows from (2.3) that B_j is a beta random variable with parameters $\sum_{i=1}^{j}\frac{(n_i-1)}{2}$ and $\frac{(n_{j+1}-1)}{2}$and that $\tilde{S}_e$, B_j are all independent random variables. Note also that the random variables $\frac{n_i S_i^2}{\sigma_i^2}$ can be expressed as

$$\frac{n_1 S_1^2}{\sigma_1^2} = \tilde{S}_e B_1 B_2 \ldots B_{k-1}, \frac{n_i S_i^2}{\sigma_i^2} = \tilde{S}_e\left(1 - B_{i-1}\right) B_i \ldots B_{k-1}$$

for $i = 2, \ldots, k-1$

$$\frac{n_k S_k^2}{\sigma_k^2} = \tilde{S}_e\left(1 - B_{k-1}\right)$$

Therefore, the generalized p value can be expressed as

$$p = 1 - E\left(H_{k-1,n-k}\left\{\frac{n-k}{k-1}\tilde{s}_b\left[\frac{n_1 s_1^2}{B_1 B_2 \ldots B_{k-1}}, \frac{n_2 s_2^2}{(1-B_1) B_2 \ldots B_{k-1}}, \ldots, \frac{n_k s_k^2}{(1-B_{k-1})}\right]\right\}\right) \tag{2.5}$$

where $H_{k-1,n-k}$is the cumulative distribution function of the F-distribution with *k-1* and *N-k* degrees of freedom. This test rejects the H_0 in (2.1) whenever $p\langle\alpha$ (Weerahandi, 1995a).

Xu-Wang's Generalized F-test

For a bigger value of k the type-I error probability of the generalized F-test exceeds the nominal level. Xu-Wang (2007a) developed some test statistics where its empirical type-I error probability does not exceed its nominal level.
Denote $v_a = \left(\mu_1, \mu_2, \ldots, \mu_{k-1}\right)'$ and $v_b = 1_{k-1}\mu_k$, where $\boldsymbol{1}_{k-1}$ is the $(k-1)\times 1$vector whose elements are all ones. Then null hypothesis in (2.1) is equal to the null hypothesis as

$$H_0 : v_a = v_b$$

The sample variances (MLEs) of the k populations are denoted by S_i^2, where

$$S_i^2 = \frac{1}{n_i}\sum_{i=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2.$$

Define

$$\begin{aligned} \bar{Y}_a &= \left(\bar{X}_1, \ldots, \bar{X}_{k-1}\right)', \quad \bar{Y}_b = 1_{k-1}\bar{X}_k, \\ S_a &= diag\left(\frac{S_1^2}{n_1 - 1}, \ldots, \frac{S_{k-1}^2}{n_{k-1} - 1}\right) S_b = \frac{1}{n_k - 1} S_k^2 1_{k-1} 1_{k-1}' \end{aligned}$$

Let ya, yb, sa and sb denote the observed values of $\bar{Y}_a$, $\bar{Y}_b$, S_a and S_b respectively. T is a generalized test variable as

$$T = Y'\left((s_a + s_b)^{-1/2}\left(diag\left(\frac{s_1^2}{U_1}, \ldots, \frac{s_{k-1}^2}{U_{k-1}}\right) + \frac{s_k^2}{U_k}\mathbf{1}_{k-1}\mathbf{1}_{k-1}'\right)(s_a + s_b)^{-1/2}\right)Y$$

and the observed value of T is given as

$$t = \left(\bar{y}_a - \bar{y}_b\right)'\left(s_a + s_b\right)^{-1}\left(\bar{y}_a - \bar{y}_b\right)$$

where $Y \sim N\left(0,\, I_{k-1}\right)$, $\quad U_i \sim \chi^2_{n_i-1}, \quad i = 1, \ldots, k.$
Under the null hypothesis in (2.1), the generalized p-value is given by

$$p = P\left(T \geq t\right)$$

and H_0 is rejected if $p < \alpha$.

The One-Stage Test

Chen and Chen (1998) developed the OS procedure since the number of samples that are required at the second stage of two-stage procedure of Bishop and Dudewicz (1981) can be large and impracticable.
For each population, the first (or any randomly chosen) n_0 $(2 \leq n_0 \leq n_i)$ observation is chosen to calculate the usual sample mean and variance as
$\bar{X}_i\left(n_0\right) = \sum_{j=1}^{n_0}\frac{X_{ij}}{n_0}$ and $S_i^2 = \sum_{j=1}^{n_0}\frac{\left(X_{ij} - \bar{X}_i(n_0)\right)^2}{(n_0 - 1)}$
Define weights U_i and V_i for the observations in the ith sample as
$U_i = \frac{1}{n_i} + \frac{1}{n_i}\sqrt{\frac{n_i - n_0}{n_0}\left(\frac{S_{[k]}^2}{S_i^2} - 1\right)}$ and $V_i = \frac{1}{n_i} - \frac{1}{n_i}\sqrt{\frac{n_0}{n_i - n_0}\left(\frac{S_{[k]}^2}{S_i^2} - 1\right)}$
Where $S_{[k]}^2$is the maximum of $(S_1^2, \ldots, S_k^2)$. Let the final weighted sample mean be defined by

$$\tilde{X}_i = \sum_{j=1}^{n} W_{ij} X_{ij}$$

where

$$W_{ij} = \begin{cases} U_i & 1 \leq j \leq n_0 \\ V_i & n_0 \langle j \leq n \end{cases}$$

Chen and Chen (1998) give the following test statistics.

$$\tilde{F}^1 = \sum_{i=1}^{k} \left(\frac{\tilde{X}_i - \tilde{X}}{S_{[k]}/\sqrt{n}} \right)^2$$

Let $\tilde{X} = \sum_{i=1}^{k} \frac{\tilde{X}_i}{k}$, $\bar{\mu} = \sum_{i=1}^{k} \frac{\mu_i}{k}$
and
$\bar{t} = \frac{\sum_{i=1}^{k} t_i}{k} = \frac{1}{k} \sum_{i=1}^{k} \frac{\tilde{X}_i - \mu_i}{S_{[k]}/\sqrt{n}} = \frac{1}{S_{[k]}/\sqrt{n}} \sum_{i=1}^{k} \frac{\tilde{X}_i - \mu_i}{k} = \frac{\tilde{X} - \bar{\mu}}{S_{[k]}/\sqrt{n}}$
Then we have

$$\tilde{F}^1 = \sum_{i=1}^{k} \left(t_i \; - \; \bar{t} \; + \; \frac{\mu_i - \mu}{S_{[k]}/\sqrt{n}} \right)^2$$

Under the null hypothesis in (2.1), it follows that $\tilde{F}^1$ is distributed as

$$Q = \sum_{i=1}^{k} (t_i - \bar{t})^2$$

which is a quadratic form in the independent student's t variates each with n_0-*1* degrees of freedom (Chen and Chen, 1998). For a given level α, and an observed value $\tilde{F}_h^1$of $\tilde{F}^1$, this test rejects the H_0 in (2.1) whenever the p-value is given as $P\left(Q_{k,n_0} > \tilde{F}_h^1\right) \; < \alpha$.

The One-Stage Range Test

In another procedure based on one stage, Chen (2001) gives the following test statistics.

$$T_1 = \frac{\tilde{X}_{\max} - \tilde{X}_{\min}}{\sqrt{z^*}}$$

where $\tilde{X}_{\max}(\tilde{X}_{\min})$ is the maximum (minumum) of $\tilde{X}_1, \cdots, \tilde{X}_k$ and z^* is the maximum of $\frac{S_1^2}{n_1}, \ldots, \frac{S_k^2}{n_k}$. Under the null hypothesis in (2.1), it follows that T_1 is distributed as

$$R = \max_{1\leq i,j\leq k} |\, t_i - t_j|$$

which is the range of k independent student's t variates each with n_0-*1* degrees of freedom. For a given level α, and an observed value t_1 of T_1, this test rejects the H_0 in (2.1) whenever the p-value is given as $P\left(R_{k,n_0-1} > t_1\right) \; < \alpha$.

A Parametric Bootstrap Approach

In the case of population variances σ_i^2s are unknown; a test statistic can be obtained by replacing σ_i^2 in (2.2) by $S_i^2$and is given by

$$\tilde{S}_b\left(S_1^2, \ldots, S_k^2\right) = \sum_{i=1}^{k} \frac{n_i \bar{X}_i^2}{S_i^2} - \frac{\left(\sum_{i=1}^{k} \frac{n_i \bar{X}_i}{S_i^2}\right)^2}{\sum_{i=1}^{k} \frac{n_i}{S_i^2}} \tag{2.6}$$

As the test statistic in (2.6) is location invariant, without loss of generality, we can take the common mean to be zero. Let $\bar{X}_{Bi} \sim N\left(0, \frac{S_i^2}{n_i}\right)$ and $S_{Bi}^2 \sim S_i^2 \chi_{n_i-1}^2 / \left(n_i - 1\right)$. Then the parametric boostrap pivot variable can be obtained by replacing $\bar{X}$, S_i^2in (2.6) by $\bar{X}_{Bi}$, S_{Bi}^2 and is given by

$$S_{bB} = \sum_{i=1}^{k} \frac{n_i}{S_{Bi}^2} \bar{X}_{Bi}^2 - \left[\sum_{i=1}^{k} \frac{n_i \bar{X}_{Bi}}{S_{Bi}^2}\right]^2 \Bigg/ \sum_{i=1}^{k} \frac{n_i}{S_{Bi}^2} \tag{2.7}$$

$\bar{X}_{Bi}$ is distributed as $Z_i \left(\frac{S_i}{\sqrt{n_i}}\right)$, where Z_i is a standard normal random variable. So the PB pivot variable in (2.7) is distributed as

$$\tilde{S}_{bB}\left(Z_i, \chi_{n_{i-1}}^2; S_i^2\right) = \sum_{i=1}^{k} \frac{Z_i^2 \left(n_i - 1\right)}{\chi_{n_i-1}^2} - \frac{\left[\sum_{i=1}^{k} \frac{\sqrt{n_i} Z_i (n_i - 1)}{S_i \chi_{n_i-1}^2}\right]^2}{\sum_{i=1}^{k} \frac{n_i (n_i-1)}{S_i^2 \chi_{n_i-1}^2}}$$

H_0 is rejected if $P\left\{\tilde{S}_{bB}\left(Z_i,\, \chi_{n_{i-1}}^2;\, s_i^2\right) \rangle\, \tilde{s}_b\right\} \; \langle\, \alpha$ (Krishnamoorthy, 2006).

3. SIMULATION STUDY

In this section we compare the CF, W, SS, BF, GF, XW, OS, OSR and PB tests according to type-1 errors and powers in different combinations of parameters and sample sizes.

3.1. **Comparison between the type I error rates of the tests.** In this section we consider the balanced and unbalanced cases from smaller to larger sample sizes where k=3 and k=5 for comparing the tests. The values for the variances vary over a large range so that $\sigma_1^2 < \ldots < \sigma_k^2$ and $\sigma_1^2 > \ldots > \sigma_k^2$. For each combination of n_i and $\sigma_i^2$the rejection rate of each testing procedure is calculated and compared with the nominal level 0.05 when the means are all equal. To estimate the type I error rates of the CF, W, SS, BF tests, we use simulation consisting of 5000 runs for each of the sample sizes and parameter configurations. CF, W, SS, BF test statistics are calculated from these generated data and type I errors are estimated by the proportion of test statistics that exceed the critical values calculated from the distributions. To estimate the type I error rates of the GF, PB, OS, OSR and XW tests, we use a two-step simulation. For estimating the type I error rates of the GF test we generate 5000 observed vectors $\left(\bar{x}_1, ..., \bar{x}_k; s_1^2, ..., s_k^2\right)$, and used 5000 runs for each observed vector to estimate the p value in (2.5). Finally the type I error rate of the GF test are estimated by the proportion of the 5000 p-values that are less than the nominal level α. The type I error rates of the PB, OS, OSR and XW tests are similarly estimated. In both cases of equal and unequal variances for k=3 and k=5 simulated type I error probabilities are given in tables 1, 2, 3 and 4.

n_i	σ_i	CF	W	SS	BF	GF	OS	OSR	PB	XW
4,4,4	(1,1,1)	.0494	.0422	.0360	.0348	.0324	.0464	.0494	.0412	.0150
	(4,4,4)	.0498	.0432	.0366	.0374	.0342	.0484	.0480	.0418	.0158
	(9,9,9)	.0492	.0388	.0366	.0338	.0332	.0462	.0466	.0372	.0132
	(1,1.25,1.5)	.0565	.0443	.0386	.0417	.0363	.0481	.0450	.0432	.0166
	(1,2,4)	.0798	.0521	.0975	.0571	.0431	.0435	.0418	.0498	.0368
	(1,4,9)	.0932	.0674	.2996	.0588	.0652	.0548	.0558	.0620	.0446
10,10,10	(1,1,1)	.0498	.0492	.0314	.0490	.0460	.0490	.0430	.0492	.0370
	(4,4,4)	.0496	.0526	.0322	.0501	.0480	.0460	.0434	.0540	.0380
	(9,9,9)	.0499	.0532	.0292	.0502	.0470	.0510	.0478	.0530	.0394
	(1,1.25,1.5)	.0548	.0522	.0376	.0506	.0482	.0502	.0508	.0532	.0426
	(1,2,4)	.0730	.0502	.1446	.0576	.0494	.0490	.0458	.0504	.0610
	(1,4,9)	.0760	.0448	.4202	.0606	.0478	.0500	.0472	.0442	.0680
30,30,30	(1,1,1)	.0474	.0490	.0224	.0474	.0466	.0470	.0460	.0484	.0532
	(4,4,4)	.0476	.0479	.0210	.0476	.0466	.0453	.0440	.0469	.0480
	(9,9,9)	.0474	.0490	.0224	.0474	.0466	.0470	.0456	.0484	.0504
	(1,1.25,1.5)	.0490	.0476	.0260	.0480	.0464	.0510	.0486	.0482	.0640
	(1,2,4)	.0694	.0499	.1481	.0645	.0493	.0458	.0446	.0505	.0708
	(1,4,9)	.0724	.0510	.4504	.0658	.0510	.0466	.0434	.0500	.0756

Table 1. Simulated type I error rates when k=3 and sample sizes are equal under nominal $\alpha = 0.05$

n_i	σ_i	CF	W	SS	BF	GF	OS	OSR	PB	XW
	(1,1,1)	.0496	.0508	.0172	.0446	.0390	.0524	.0514	.0524	.0166
	(4,4,4)	.0476	.0510	.0155	.0452	.0396	.0500	.0488	.0516	.0160
3,5,7	(9,9,9)	.0485	.0504	.0144	.0396	.0342	.0488	.0464	.0508	.0178
	(1,1.25,1.5)	.0336	.0500	.0174	.0410	.0322	.0468	.0450	.0486	.0166
	(1,2,4)	.0292	.0512	.0264	.0264	.0406	.0468	.0456	.0494	.0248
	(1,4,9)	.0332	.0566	.0350	.0638	.0486	.0468	.0444	.0542	.0452
	(1.5,1.25,1)	.0792	.0586	.0214	.0500	.0448	.0522	.0510	.0592	.0266
	(4,2,1)	.1852	.0680	.0908	.0712	.0568	.0524	.0506	.0676	.0368
	(9,4,1)	.2332	.0766	.3554	.0728	.0734	.0502	.0484	.0646	.0432
	(1,1,1)	.0486	.0492	.0302	.0454	.0448	.0502	.0482	.0498	.0158
	(4,4,4)	.0496	.0504	.0298	.0408	.0452	.0536	00502	.0502	.0358
	(9,9,9)	.0490	.0490	.0296	.0470	.0454	.0454	.0430	.0488	.0396
7,10,13	(1,1.25,1.5)	.0388	.0460	.0344	.0476	.0426	.0502	.0474	.0460	.0142
	(1,2,4)	.0300	.0500	.1284	.0570	.0446	.0478	.0476	.0488	.0560
	(1,4,9)	.0314	.0534	.3746	.0632	.0524	.0484	.0472	.0502	.0640
	(1.5,1.25,1)	.0816	.0560	.0364	.0580	.0388	.0540	.0502	.0524	.0390
	(4,2,1)	.1462	.0540	.1360	.0632	.0526	.0516	.0504	.0588	.0470
	(9,4,1)	.1688	.0548	.4136	.0666	.0580	.0512	.0516	.0510	.0534
	(1,1,1)	.0494	.0534	.0254	.0540	.0506	.0502	.0492	.0522	.0500
	(4,4,4)	.0505	.0460	.0226	.0474	.0442	.0476	.0456	.0452	.0506
	(9,9,9)	.0496	.0478	.0224	.0468	.0462	.0456	.0472	.0470	.0536
20,25,30	(1,1.25,1.5)	.0418	.0528	.0314	.0540	.0502	.0494	.0506	.0514	.0570
	(1,2,4)	.0386	.0470	.1370	.0598	.0458	.0484	.0466	.0456	.0672
	(1,4,9)	.0394	.0470	.4254	.0634	.0492	.0472	.0470	.0482	.0684
	(1.5,1.25,1)	.0694	.0521	.0305	.0559	.0509	.0507	.0488	.0502	.0478
	(4,2,1)	.1110	.0482	.1446	.0652	.0486	.0542	.0488	.0468	.0454
	(9,4,1)	.1248	.0464	.4456	.0678	.0484	.0520	.0484	.0454	.0470

Table 2. Simulated type I error rates when k=3 and sample sizes are unequal under nominal α=0.05

n_i	σ_i	CF	W	SS	BF	GF	OS	OSR	PB	XW
4,4,4,4,4	(1,1,1,1,1)	.0486	.0472	.0494	.0334	.0618	.0484	.0480	.0322	.0162
	(4,4,4,4,4)	.0480	.0472	.0472	.0344	.0568	.0530	.0538	.0356	.0120
	(9,9,9,9,9)	.0462	.0522	.0510	.0318	.0640	.0510	.0516	.0358	.0108
	(1,1.25,1.5,1.75,2)	.0638	.0634	.0640	.0442	.0740	.0494	.0494	.0470	.0226
	(1,2,4,6,8)	.0920	.0804	.1899	.0571	.0936	.0465	.0467	.0528	.0342
	(1,4,9,13,18)	.0978	.0856	.4198	.0556	.1092	.0538	.0532	.0580	.0358
10,10,10,10,10	(1,1,1,1,1)	.0486	.0518	.0420	.0468	.0586	.0516	.0460	.0496	.0522
	(4,4,4,4,4)	.0488	.0520	.0404	.0470	.0580	.0492	.0434	.0494	.0484
	(9,9,9,9,9)	.0498	.0548	.0436	.0488	.0616	.0498	.0442	.0506	.0508
	(1,1.25,1.5,1.75,2)	.0626	.0518	.0494	.0570	.0610	.0512	.0518	.0488	.0586
	(1,2,4,6,8)	.0852	.0514	.1844	.0566	.0644	.0512	.0518	.0488	.0760
	(1,4,9,13,18)	.0880	.0516	.5476	.0664	.0684	.0496	.0490	.0490	.0768
30,30,30,30,30	(1,1,1,1,1)	.0494	.0484	.0300	.0490	.0510	.0476	.0504	.0484	.0666
	(4,4,4,4,4)	.0494	.0500	.0310	.0490	.0538	.0468	.0472	.0504	.0736
	(9,9,9,9,9)	.0504	.0550	.0318	.0496	.0570	.0532	.0528	.0548	.0688
	(1,1.25,1.5,1.75,2)	.0582	.0484	.0376	.0562	.0516	.0490	.0476	.0488	.0832
	(1,2,4,6,8)	.0830	.0494	.2822	.0790	.0542	.0474	.0462	.0470	.1044
	(1,4,9,13,18)	.0854	.0502	.5812	.0814	.0502	.0538	.0536	.0486	.1000

Table 3. Simulated type I error rates when k=5 and sample sizes are equal under nominal α=0.05

n_i	σ_i	CF	W	SS	BF	GF	OS	OSR	PB	XW
	(1,1,1,1,1)	.0466	.0592	.0326	.0378	.0656	.0452	.0442	.0462	.0132
	(4,4,4,4,4)	.0500	.0630	.0334	.0394	.0686	.0468	.0454	.0470	.0164
	(9,9,9,9,9)	.0504	.0636	.0336	.0410	.0666	.0462	.0450	.0496	.0108
3,4,5,6,7	(1,1.25,1.5,1.75,2)	.0356	.0548	.0330	.0454	.0650	.0490	.0486	.0402	.0210
	(1,2,4,6,8)	.0380	.0606	.0616	.0650	.0688	.0460	.0454	.0480	.0486
	(1,4,9,13,18)	.0326	.0682	.0720	.0620	.0790	.0482	.0460	.0504	.0536
	(2,1.75,1.5,1.25,1)	.1088	.0804	.0426	.0554	.0884	.0526	.0512	.0600	.0168
	(8,6,4,2,1)	.2114	.0862	.2180	.0580	.1000	.0516	.0502	.0590	.0404
	(18,13,9,4,1)	.2236	.0920	.5140	.0618	.1124	.0560	.0540	.0582	.0492
	(1,1,1,1,1)	.0506	.0521	.0420	.0462	.0608	.0494	.0474	.0510	.0470
	(4,4,4,4,4)	.0508	.0530	.0430	.0494	.0622	.0556	.0524	.0534	.0470
	(9,9,9,9,9)	.0504	.0535	.0466	.0496	.0672	.0528	.0496	.0570	.0474
7,9,11,13,15	(1,1.25,1.5,1.75,2)	.0378	.0522	.0518	.0596	.0584	.0496	.0490	.0496	.0588
	(1,2,4,6,8)	.0426	.0528	.2398	.0770	.0602	.0484	.0460	.0514	.0860
	(1,4,9,13,18)	.0448	.0580	.4958	.0804	.0702	.0504	.0490	.0540	.0822
	(2,1.75,1.5,1.25,1)	.0926	.0526	.0488	.0522	.0638	.0516	.0510	.0484	.0442
	(8,6,4,2,1)	.1744	.0612	.2820	.0650	.0724	.0508	.0514	.0472	.0460
	(18,13,9,4,1)	.1832	.0588	.5740	.0672	.0732	.0512	.0476	.0486	.0526
	(1,1,1,1,1)	.0486	.0564	.0304	.0494	.0590	.0564	.0512	.0564	.0570
	(4,4,4,4,4)	.0474	.0478	.0310	.0490	.0510	.0520	.0470	.0476	.0622
	(9,9,9,9,9)	.0510	.0542	.0324	.0516	.0564	.0540	.0464	.0532	.0668
20,23,26,29,32	(1,1.25,1.5,1.75,2)	.0438	.0506	.0390	.0568	.0524	.0536	.0512	.0492	.0834
	(1,2,4,6,8)	.0466	.0436	.2486	.0726	.0472	.0532	.0466	.0448	.0982
	(1,4,9,13,18)	.0496	.0502	.5488	.0790	.0556	.0528	.0496	.0494	.0918
	(2,1.75,1.5,1.25,1)	.0778	.0478	.0448	.0558	.0582	.0422	.0422	.0476	.0482
	(8,6,4,2,1)	.1360	.0466	.2870	.0736	.0518	.0488	.0468	.0560	.0504
	(18,13,9,4,1)	.1310	.0494	.5864	.0778	.0638	.0496	.0476	.0558	.0504

Table 4. Simulated type I error rates when k=5 and sample sizes are unequal under nominal α=0.05

We observe the following from the numerical results in Tables 1, 2, 3 and 4. The CF and SS tests seem to have a type I error probability exceeding the nominal level for the balanced case and small sample sizes. In the case of extreme heteroscedasticity the W, BF, GF and PB tests exceed the nominal level. However, the OS, OSR and XW tests are superior to the other tests. The W, GF and PB tests also seem to be very conservative, when the sample sizes are large. The CF, SS and BF tests exceed the nominal level when the sample sizes are proportional to variances for small sample sizes. The W, GF, OS, OSR, PB tests seem to be very conservative not only for the small sample sizes but also for the large samples. However, the XW test exceeds the nominal level for the large sample sizes. The CF, W, BF, SS and GF tests exceed the nominal level when variances and sample sizes are inversly

proportional. However, the OS, OSR and XW tests seem to be very conservative. The W, GF and PB tests have similarly results when the sample sizes are large. For a bigger value of k the CF, W, SS, BF, GF tests exceed the nominal level when the sample sizes are small. The SS, BF and XW tests have similar results when the sample sizes are large. The OS, OSR and PB tests seem to be very conservative not only for the small sample sizes but also for the large sample sizes. For all cases similar results were found. It appears that the PB, OS and OSR tests are superior to the other tests.

3.2. **Comparison Between The Powers Of The Tests.** For each combination of n_i and $\sigma_i^2$the rejection rate of each testing procedure is calculated and compared with the nominal level 0.05 when the means are not all equal. In this section we use 5000 runs for each of the sample sizes and parameter configurations to alculate the powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests. For k=3 and k=5 we provide the powers of these tests.

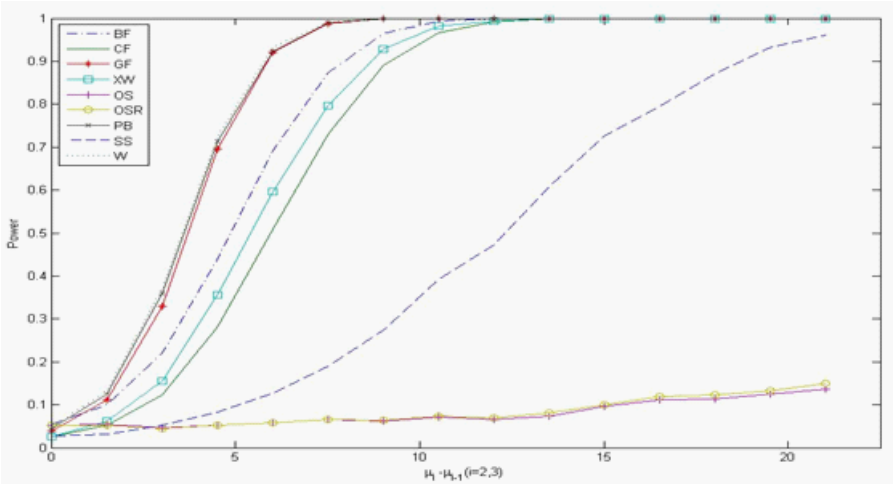

Figure 1. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests under nominal α=0.05 for k=3, n=3, 5, 7 and σ_i^2=1, 4, 9

Figure 2. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests under nominal α=0.05 for k=3, n=3, 5, 7 and σ_i^2=9, 4, 1

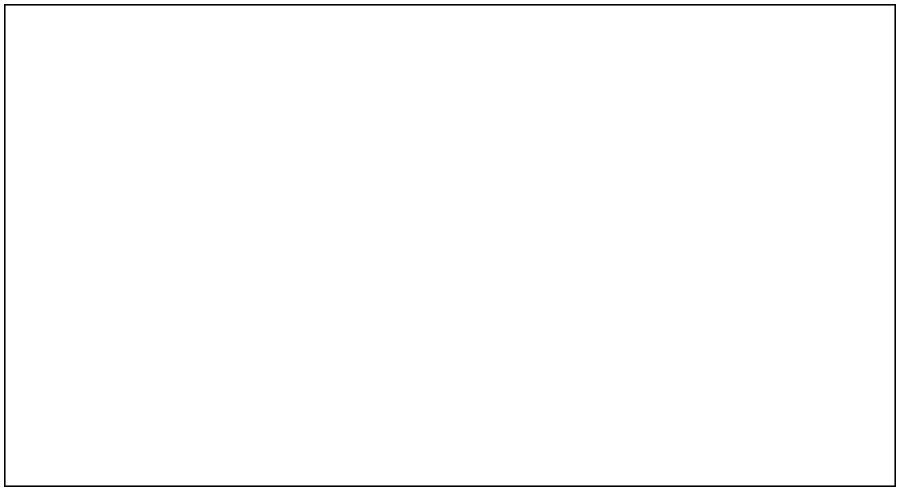

Figure 3. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests under nominal α=0.05 for k=5, n=3, 4, 5, 6, 7 and σ_i^2=1, 4, 9, 13, 18

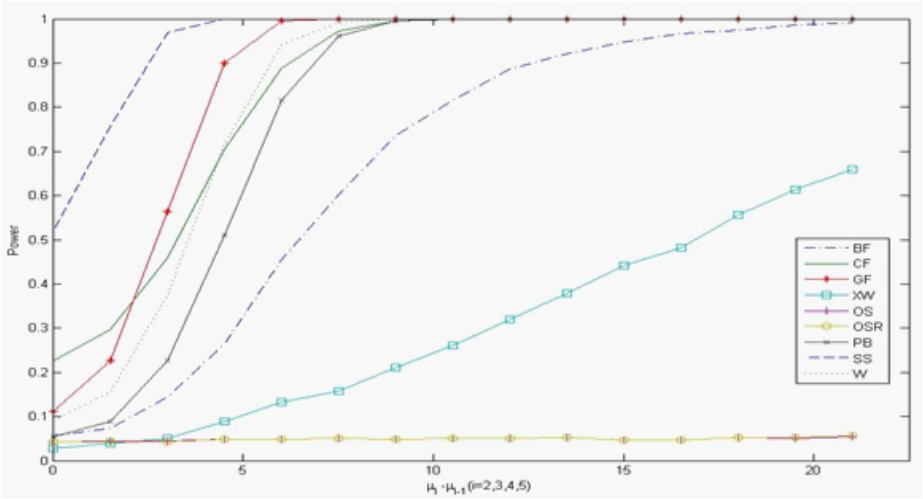

Figure 4. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests under nominal α=0.05 for k=5, n=3, 4, 5, 6, 7 and σ_i^2=18, 13, 9, 4, 1

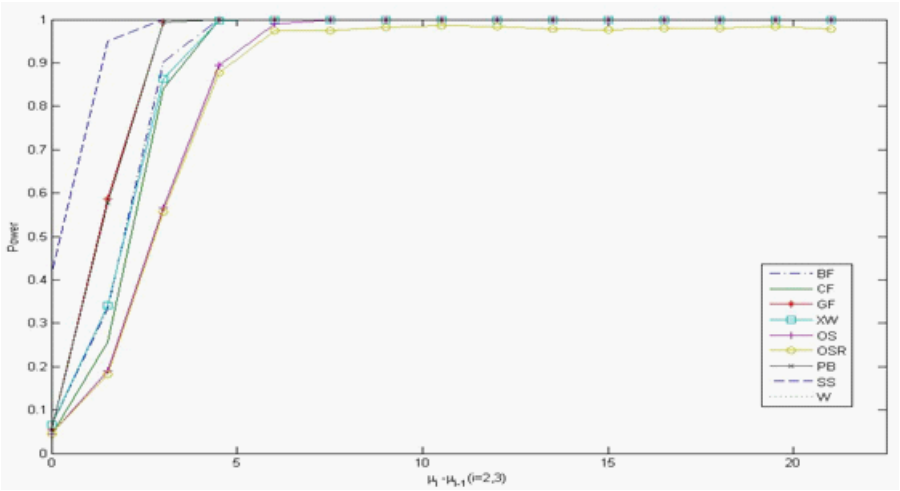

Figure 5. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests for k=3, n=20, 25, 30 and σ_i^2=1, 4, 9

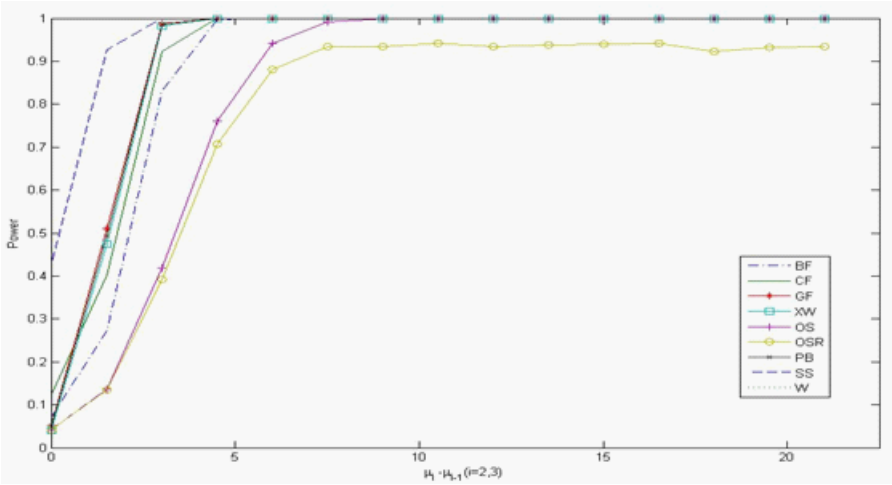

Figure 6. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests for k=3, n=20, 25, 30 and σ_i^2=9, 4, 1

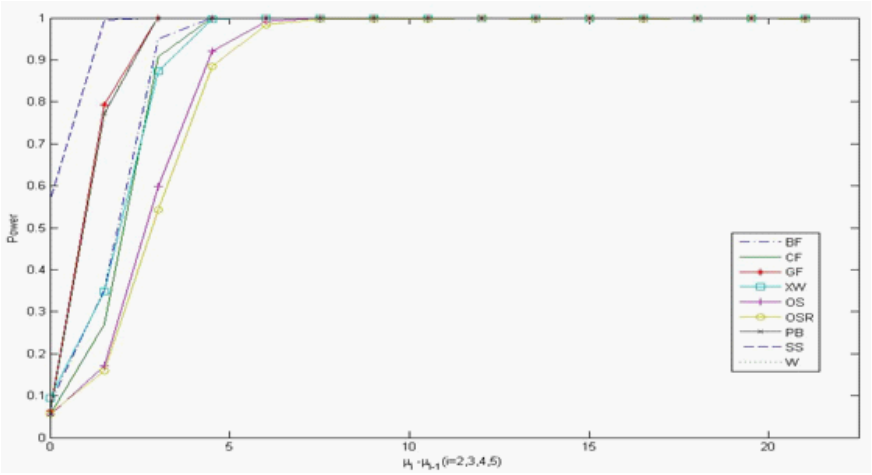

Figure 7. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests for k=5, n=20, 23, 26, 29,32 and σ_i^2=1, 4, 9, 13,18

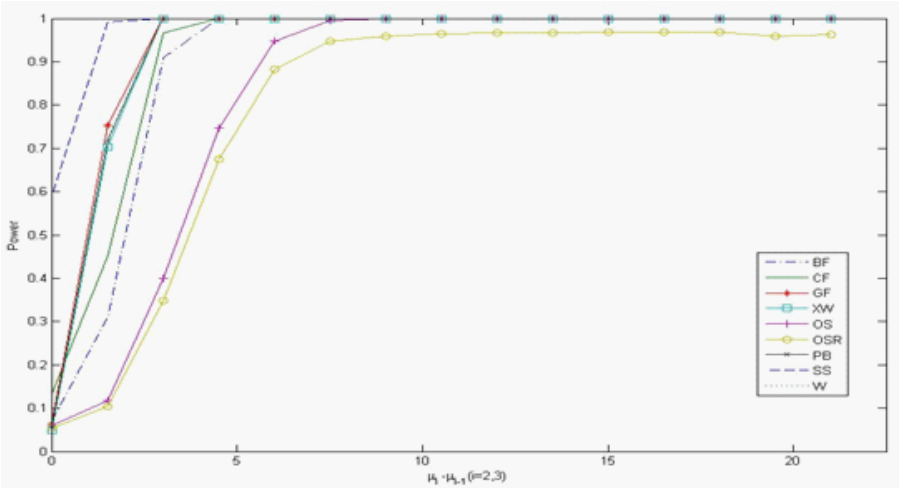

Figure 8. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests for k=5, n=20, 23, 26, 29,32 and σ_i^2=18, 13, 9, 4, 1

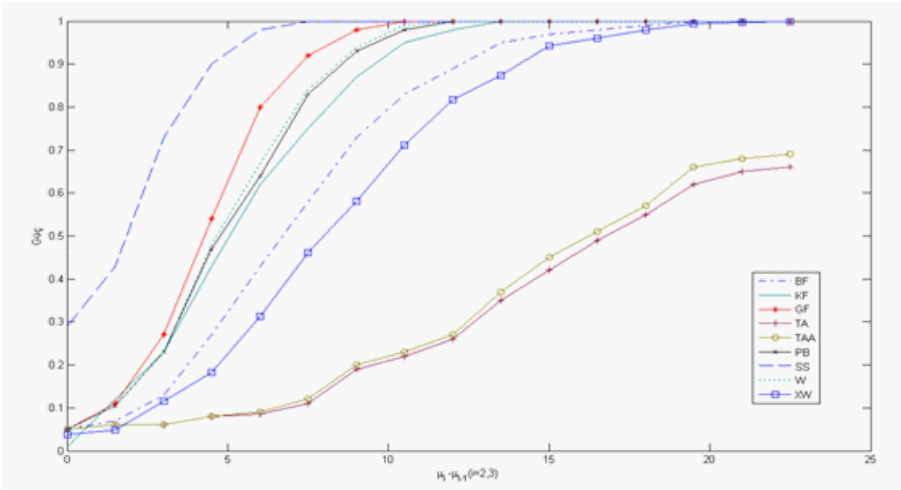

Figure 9. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests under nominal α=0.05 for k=3, n=4, 4, 4 and σ_i^2=1, 4, 9

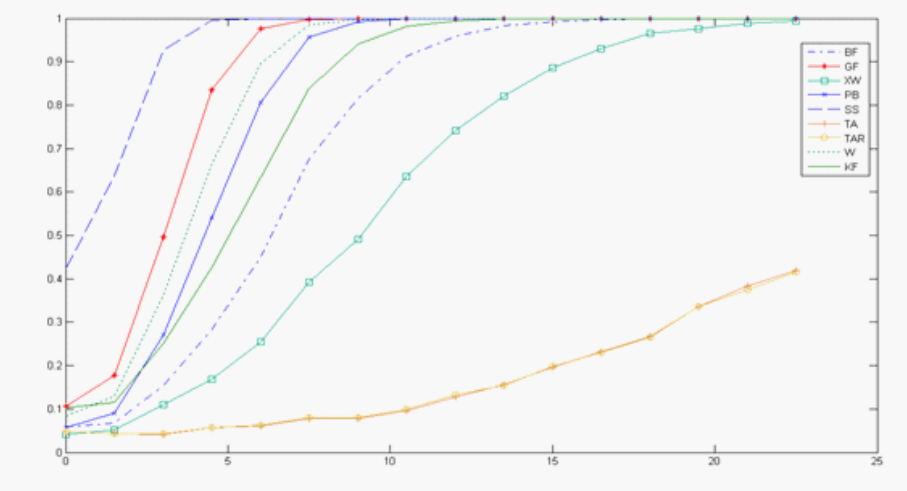

Figure 10. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests under nominal α=0.05 for k=3, n=30, 30, 30 and σ_i^2=1, 4, 9

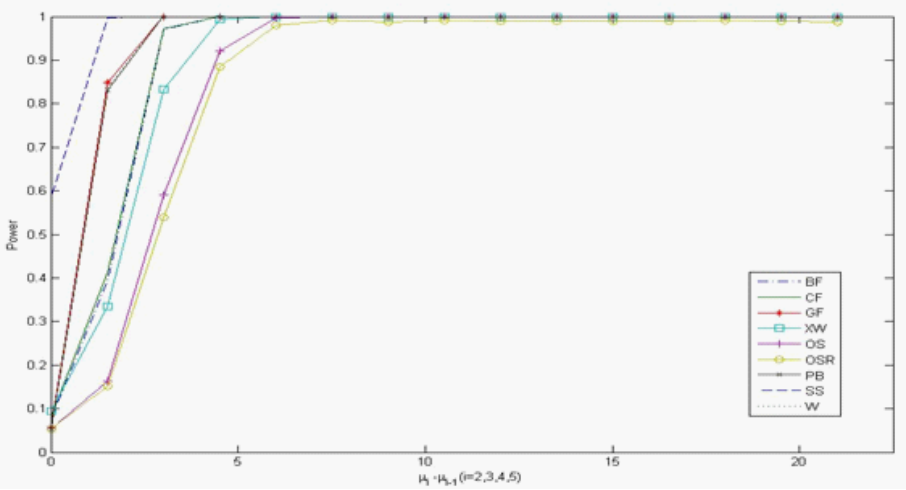

Figure 11. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests under nominal α=0.05 for k=5, n=4, 4, 4, 4, 4 and σ_i^2=1, 4, 9, 13, 18

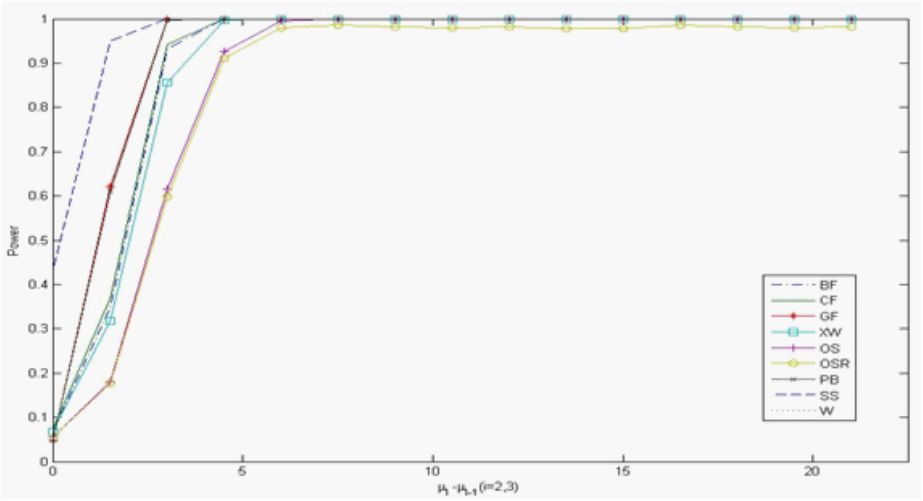

Figure 12. Powers of the CF, W, SS, BF, OS, OSR, GF, PB, XW tests under nominal α=0.05 for k=5, n=30, 30, 30, 30, 30 and σ_i^2=1, 4, 9, 13, 18

We once again observe from these figures that the tests control the type I errors. The OS and OSR tests are badly affected, especially with small and different sample sizes. These tests appear to be less powerful than the other tests although their type I error rates are close to the intended level 0.05. In most cases the SS test is disregarded because of its type I error rates exceeding the intended level 0.05. The other tests exhibit close power proporties provided the type I error rates are close to the intended level 0.05. Powers of tests become different from each other when the variances and sample sizes are inversely proportional for small sample sizes. These differences increase, especially for bigger values of k. In most cases the GF, W and PB tests appear to be more powerful than the other tests. In particular, the GF test is superior to the other tests, except for small sample sizes and bigger values of k, because its type I error rates exceed the intended level 0.05. In this case the PB test is superior to the other tests.

4. CONCLUSION

In this simulation study for a range of choices of sample sizes and parameter configurations we compared the performance of the above tests for testing the equality of means of one-way ANOVA models under heteroscedasticity. The CF test is not an appropriate test for heteroscedasticity because its type I error rates exceed the intended level 0.05. The same is true for the SS test. The OS and OSR tests appear to be less powerful than the other tests even though their type I error rates

are close to the intended level 0.05, regardless of the sample sizes, value of the error variances and the number of means being compared.

The W and PB and especially the GF tests appear to be more powerful than the other tests when k=3 and the sample sizes are small (n_1, n_2, n_3=3, 5, 7). The W and PB tests are superior to the other tests when k=5 and the sample sizes are small (n_1, n_2, n_3=3, 5, 7). When the sample sizes are large the GF, W and PB tests are more powerful than the other tests when both k=3 and k=5. In this case the XW test is also powerful when the variances and sample sizes are inversely proportional.

Although the empirical type I errors of the tests based on the OS procedure are close to their nominal level, the powers of these tests are not as high as those of the GF, W and PB tests. For this reason, the GF, W and PB tests can be used instead of tests based on the OS procedure.

ÖZET: İkiden fazla yığının ortalamalarının eşitliği hipotezinin testi amacıyla kullanılan klasik F testi, normallik ve yığın varyanslarının homojenlik varsayımına dayanır. Bu varsayımlar özellikle varyansların homojenlik varsayımı sağlanmadığında klasik F testinin kullanılması uygun olmamaktadır. Bu durum özellikle örneklem hacmi büyük olmadığında, önemli bir sıkıntı doğurmaktadır. Literatürde bu konuyla ilgili bir çok test istatistiği geliştirilmiştir. Bu çalışmada Brown-Forsythe, Weerahandi'nin Genelleştirilmiş F, Parametrik Bootstrap, Scott-Smith, One-Stage, One-Stage Range, Welch ve Xu-Wang testleri tanıtılmış ve testlerin farklı yığın parametreleri ve örnek hacimleri altında deneysel I.tip hata oranı ve testin gücü bakımından karşılaştırılması yapılmıştır.

References

[1] Bishop, T.A. and Dudewicz, E.J. Heteroscedastic ANOVA, Sankhya **43**B:40-57 (1981).

[2] Brown, M.B., Forsythe, A.B. The small sample behavior of some statistics which test the equality of several means, Technometrics **16**: 129-132 (1974).

[3] Chen, S. and Chen, J.H. Single-Stage Analysis of Variance Under Heteroscedasticity, Communications in Statistics Simulations **27**(**??**): 641-666 (1998).

[4] Chen, S. One stage and two stage statistical inference under heteroscedasticity, Communications in Statistics Simulations **30**(**??**): 991-1009 (2001).

[5] Krishnamoorthy, K., Lu, F., Thomas, M. A parametric boostrap approach for ANOVA with unequal variances: fixed and random models, Computational Statistics and Data Analysis, **51**:5731-5742 (2006).

[6] Weerahandi, S., ANOVA under unequal error variances, Biometrica, **38**:330-336 (1995a).

[7] Weerahandi, S., Exact statistical method for data analysis, Springer-Verlag, New York, 2-50 (1995).

[8] Weerahandi, S., Generalized inference in repeated measures: Exact methods in MANOVA and mixed models, Wiley, New York, 1-60 (2004).
[9] Welch, B.L., The generalization of student'sproblem when several different population variances are involved, Biometrika,3 4:28-35(1947).
[10] Welch, B.L., On the comparison of several mean values: An alternative approach, Biometrica, **38:**330-336 (1951).
[11] Scott, A.J. ve Smith, T.M.F., Interval Estimates for Linear Combinations of Means, Applied Statistics, **20**(**??**):276-285 (1971).
[12] Tsui, K. and Weerahandi, S., Generalized p-Values in Significance Testing of Hypotheses in the Presence of Nuisance Parametres, Journal of the American Statistical Assocation, **84**:602-607 (1989).
[13] Xu, L. and Wang, S. A new generalized p-value for ANOVA under heteroscedasticity, Statistics and Probability Letters, **78**:963-969 (2007a).
[14] Xu, L. and Wang, S. A new generalized p-value and its upper bound for ANOVA under unequal erros variances, Communications in Statistics Theory and Methods, **37:**1002-1010 (2007b).

Current address: Gazi University Faculty of Science and Art Depermant of Statistics Teknikokullar Ankara

E-mail address: eyigit@gazi.edu.tr; fikri@gazi.edu.tr

URL: http://communications.science.ankara.edu.tr