A Robust Likelihood Ratio Test for Testing Equal Means in the Presence of Unequal Variance

Achut Adhikari
University of Northern Colorado, Department of Applied Statistics and Research Methods, McKee 518, Campus Box 124, Greeley, CO 80639

Abstract

Testing equality of several means is a common task for consulting statisticians. The usual assumptions for Fisher's *F* test require *k* independent random samples from *k* normal distributions with equal variances. In this work, I study the sampling distribution of the likelihood ratio statistic, denoted λ, within the context of unequal variances. My searches for a more general test are in response to the elevated alpha of the *F* test, in the presence of unequal variances. This work is valuable because the empirical alphas for the *F* test and the -2log (λ) are often higher than the intended α = 0.05, in the presence of unequal variance, especially for small samples. I present a framework for a robust test which relies on the sampling distribution of λ and not the transformed λ. Hence, we do not have to depend on the chi-square distribution for an approximation.

The focus of the study is as follows: First, generate the critical values for the robust test (5th percentile of the likelihood ratio statistic, denoted P05) when sampling from 4 normal distributions/populations. The P05 was generated under a variety of variance patterns. The hope is that for a fixed value of *n* (replications per treatment level), with equal means, the unequal variance will not greatly affect the empirical rejection rates.

Simulation results revealed for a fixed *n* (such as *n* = 3) the empirical alpha for the *F* test began very close to 0.05 (when all four variances were equal) and increased up to 0.0622 (when all four variances were not equal). As *n* increases to 15, the empirical alpha for first variance pattern (four variances were equal) was around 0.05 and hovered in between 0.056 – 0.058 for other variance patterns (all four variances were not equal). The similar trend was exhibited for the -2 log (λ); however the alphas start much higher for this test. For example, the alpha for *n*=3 with equal variance was 0.189.

Finally, simulations revealed for each value of *n*, the P05 values are very stable under the effect of different variance patterns. For example, P05 hovers around 0.0022 when *n* = 3, regardless of variance pattern; for *n* = 5 the P05 value hovers around 0.0056; for *n* = 7 the value hovers around 0.0086. So indeed the 5th percentile values are stable (within a fixed value of n) over the variance patterns we studied. It remains to be seen if these P05 values serve as effective critical values for a test of equal means. This question will be studied in subsequent research.

Key Words: Fisher's *F* test, chi-square approximation, Likelihood Ratio Test, α

1. Introduction

Testing equality of means is a very common task encountered by researchers and statistical consultants. To test equal means using Fisher's *F* test, the analysis of variance requires the usual assumptions of *k* independent random samples from *k* normal distributions with equal variances (Kuehl, 2000).

In this work I studied the sampling distribution of the likelihood ratio test statistic, λ, in testing equality of four means. From these sampling distribution results, I will develop a test based on the 5th percentile of λ's, which does not depend on the Chi-Square approximation. In years past, when this approximation was first developed (Wilks, 1938), statisticians had to depend on an approximation, because cheap and high speed personal computers were not easily available at that time. They would make their decision regarding the rejection of Ho based on a conveniently available Chi-Square table.

In addition to the aforementioned sampling distribution, I studied the rejection rates for the conventional *F* test and the -2log (λ) in testing equality of means. Many studies have examined the robust properties of the *F* test under minor to moderate violations of assumptions (Box, 1954, Horsnell, 1953), and I wanted to observe these effects in the exact context of my sample sizes and parameter settings.

In summary, my main goal is to derive the likelihood ratio test, and study the distribution of the λ statistic without using the chi-square approximation. From this sampling distribution, we can obtain critical values for a robust test (5th percentile), and tabulate an array of these values to be used by consultants. Using SAS simulations, a table of the 5th percentiles was generated under the effect of different variance patterns.

If the fifth percentile (PO5) is stable over different variance patterns (for fixed *n*, fixed number of populations and equal means), consulting statisticians *may* be able to test equality of means, for small *n*, without a concern about how variances differ.

2. Review of Literature

2.1 Likelihood Ratio Test

> The notion of using the magnitude of the ratio of two probability density functions as the basis of a best test or of a uniformly most powerful test can be modified and made intuitively appealing, to provide a method of constructing a test of a composite hypothesis against an alternative composite hypothesis or of constructing a test of a simple hypothesis against an alternative composite hypothesis, when a uniformly most powerful test does not exist. This method leads to tests called likelihood ratio tests. (Hogg and Craig, 1995, P. 413)

The likelihood ratio test (in the context of my study) is based on the assumption of normally distributed random errors and four independent random samples. The ratio of likelihood function (when Ho is true) to the likelihood function (when Ha is true) is called the likelihood ratio and is defined by λ (x_1, x_2, …, x_n) = $\frac{L(\widehat{w})}{L(\widehat{\Omega})}$.The values of λ ranges from 0 to 1. A small value of λ leads to rejection of Ho. Intuitively, if the numerator is small in relation to the denominator, this indicates that the data being tested is not likely to originate from conditions existing if the null hypothesis is true.

2.2 An Application of the of Likelihood Ratio Test (LRT)

Allison et al. (1999) tested the robustness of the Likelihood Ratio Test in a variance component Quantitative-trait Loci-Mapping procedure. They investigated the robustness of the LRT for a variance component quantitative trait locus detection procedure to violations of normality. A FORTAN program was written to simulate data for the various conditions to be tested where they considered various type I error rates and distributions.

Authors argued that some types of non-normality produced type I error rates in excess of the normality whereas others did not; and that the degree of type I error –rate inflation appeared to be directly related to the residual sibling correlation. In the normal data, the type I error rates were found consistent with the nominal alpha levels. In some mixture data the empirical type I error rates were consistent with the nominal alpha levels whereas in others the type I error rates was slightly exceeded the nominal alpha levels. The excess error probability was found higher for small samples as compared with the larger sample size. In the chi-square distribution, the empirical type I error rates were exceed the nominal alpha levels by a degree that is directly related to the degree of correlation among the sibling phenotypes. Result for the symmetric but kurtotic Laplace distributions were found similar to those for chi square distributions. They observed that, for mixture distributions, increased sample size considerably decreased the size of the test whereas the highly kurtotic distributions showed virtually no improvement with increased sample size.

2.3 The generalized Likelihood Ratio test for the Behrens-Fisher problem

> A solution is suggested for the Behrens-Fisher problem of testing the equality of the means from two normal populations where variances are unknown and not assumed equal, by considering an adaption of the generalized likelihood ratio test. The test developed and called the adjusted likelihood ratio test has size close to the nominal significance level and compares favorably with regard to size and power to the Welch-Aspin test. An asymptotic result showed the connection between the generalized likelihood ratio test and the most commonly used test statistic for the Behrens-Fisher problem. (Cox & Jaber, 1990, P. 63)

They defined and established the relationship between Likelihood ratio test (LRT) and the Behrens-Fisher distribution. Behrens-Fisher distribution was defined by the following quantity:

$$V = \frac{(\mu 2 - \mu 1) - (\overline{y2} - \overline{y1})}{\left(\frac{s1^2}{n1} + \frac{s2^2}{n2}\right)^{1/2}}$$

Where, $\mu_{i,}$ $\bar{y}_i$, $S^2{}_i$ and n_i ($i = 1, 2$) are population means, sample means, sample variance and sample size respectively. Similarly, they defined the generalized likelihood ratio test for two populations as:

$$\Lambda = \frac{\underset{H1}{\sup(likelihood)}}{\underset{H0}{\sup(likelihood)}}$$

When $\sigma^2{}_1 = \sigma^2{}_2$ the LRT is simply the two sample t-test but when $\sigma^2{}_1 \neq \sigma^2{}_2$, the LRT is given by:

$$\Lambda = \left(\frac{\widehat{\sigma 1}^2}{s1^2}\right)^{1/2n1} \left(\frac{\widehat{\sigma 2}^2}{s2^2}\right)^{1/2n2}$$

They demonstrated that, the asymptotic distribution of λ is obviously Chi-square distribution with one degree of freedom ($\lambda = V^2$). The LRT was very conservative and allowed the size to be much smaller than α for certain values of C. The adjusted likelihood ratio test was in contrast to the method of Bozdogan and Ramirez who adapted the generalized likelihood ratio test by adjusting the degrees of freedom in the approximate chi-square distribution of λ. They estimated size at various values of C for all the tests, the estimate being the proportion of number of times that the null hypothesis of equal population means was rejected.

Authors showed that the size of Cochran's test and McCullough-Banerjee's test was always much less than the nominal significance level, the later always had slightly less size than the formal. LRT, Welch, Dixon and Massey, and Welch and Aspin tests lied within the sampling confidence limits of α. It was demonstrated that LRT and the Welch-Aspin test were comparable and performed well regarding control of size. It was found that the Cochran test and McCullough-Banerjee test always had size less than the nominal significance level, while the Dixon & Massey test nearly always had size greater than the sizes of all the other tests. These three tests had the largest deviations from the nominal significance level. The Welch test performed better than these three tests but not as well as LRT and the Welch-Aspin test. LRT and the Welch-Aspin test had comparable power, although in most cases the power of LRT was just slightly higher than that of the Welch-Aspin test.

A comparison of the Bozdogan and Ramirez method for using the LRT for the Behrens-Fisher problem was made with LRT. It was found that the adjusted likelihood ratio test described in this research controlled the size better than that of Bozdogan and Ramirez (BR). It was also found that the size of BR was always slightly greater than that of LRT. Simulation studies of the power showed BR to have power equal to LRT or marginally greater. This showed that LRT and BR were very similar to behavior but LRT keeping better to nominal significance level.

In brief, the authors recommended the Welch-Aspin test in terms of size and power to test equal means. The proposed generalized likelihood ratio test was shown to be comparable to the Welch-Aspin test and in most instances will be slightly more powerful.

2.4 A Robust Likelihood Ratio Test: Testing Equal Means in the Presence of Unequal Variance

Yan (2009) studied the sampling distribution of the likelihood ratio statistic, denoted by λ, within the context of unequal variances. Again, simulations were built upon independent random sampling from four normal distributions (with unequal variance). The variance patterns in this study were 1,1,1, σ_4^2 , where σ_4^2 = 1,2,4,6,8,10,15, and 30 . Empirical alpha values for the *F* test, and the Chi-square approximation were studied, and the P05 values were generated.

Simulation results revealed for a fixed *n* (such as *n* = 3) the empirical alpha for the *F* test began very close to 0.05 and increased as the variance for population four was increased to 30. For example (*n*=3) the alpha progressed from 0.051 to 0.143. As *n* increased to 25, the final alpha hovered around 0.10. The same increasing trend was exhibited for the -2 log (λ); however the alphas started much higher for this test. For example, the alpha for equal variance and *n* = 3 was 0.188.

Yan also used these P05 values as critical values for test of equal means in a subsequent set of simulations. The new test seemed to maintain the prescribed alpha (0.05) in response to different variance patterns (changes in variance four). However it should be noted that the success of these simulations rested upon a circular logic; the variance patterns used in developing the table of critical values were the same pattern used to generate the data. The utility of the new test should be evaluated with data generated from a set of more general variance patterns.

Yan’s work is the direct precursor for my current work. In another words, my current work is the extension and continuation of Yan’s work. The parameter settings on my study are much difference than the Yan’s work which is presented on methods section.

3. Methods

3.1 Simulation details

To study the sampling distribution of λ and the empirical rejection rates of *F* and -2 log (λ), simulations were performed using Proc IML (SAS: Interactive Matrix Language, 2004). Alpha was set at 0.05 and there were 20,000 trials for each sample size-parameter combination. Each trial was composed of four independent random samples from four normal distributions with parameter combinations as outlined below. Within each simulation, the rejection rate of the *F* test (reject Ho: yes or no) and the rejection rate for -2 log (λ) was counted. Also the λ values were stored for the calculation of P05 and production of graphs. Univariate procedure using SAS (9.2) was used to calculate the 5th percentile (P05) of λ.

3.2 Parameter settings

Random samples were drawn from four normal distributions with equal means but the variances were unequal. Sample sizes ranged from 3 to 15(increments of 2, 3 and 5). Emphasis was given on examining small sample sizes. Population means were all set equal to zero. For each sample size studied, there were seven different variance patterns. Only the first variance pattern meets the assumptions required for Fisher’s *F* test. So, for each sample size, there were 7 combinations of these four variances which are shown in table 1.

Table 1: Variance Patterns

Parameters	Variance Codes						
	1	2	3	4	5	6	7
σ_1^2	1	1	0.25	0.5	2	2.5	4
$\sigma 2^2$	1	2	0.50	1.0	4	5	8
σ_3^2	1	3	0.75	1.5	6	7.5	12
σ_4^2	1	4	1.00	2.0	8	10	15

Sample sizes: 3, 5, 7, 10, 15

3.3 Derivation of Likelihood ratio test (LRT)

In the context of my study, I sampled from four populations which all were normally distributed and I wanted to test the following hypotheses:

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, variances are unspecified.
Ha: μ_i were not equal, variances are unspecified.

3.4 Maximum Likelihood Estimator (MLE) under Ho

Under Ho, we had five equations and five unknowns. We have the following four populations with equal means (μ) and differing variances (θ_i). We took n random samples from each of the populations such that:

$x_1, x_2, \ldots, x_n$ was a random sample from a population, where X ~N(μ, θ_1),
$y_1, y_2, \ldots, y_n$ was a random sample from a population, where Y ~N(μ, θ_2),
$z_1, z_2, \ldots, z_n$ was a random sample from a population, where Z ~N(μ, θ_3),
$w_1, w_2, \ldots, w_n$ was a random sample from a population, where W ~N(μ, θ_4).

Under null hypothesis, Ho true implies that the four population means are equal and this lets us use μ instead of $\mu_{i.}$.Then the likelihood function is the joint pdf of four populations as a function of μ and θ_i , which can be shown as follows:

$$L(\boldsymbol{\mu}, \theta_1, \theta_2, \theta_3, \theta_4)= \prod_{i=1}^{n} f(x_i | \boldsymbol{\mu},\theta_1)\prod_{i=1}^{n} f(y_i | \boldsymbol{\mu},\theta_2)\prod_{i=1}^{n} f(z_i | \boldsymbol{\mu},\theta_3)\prod_{i=1}^{n} f(w_i | \boldsymbol{\mu},\theta_4) \quad =$$

$$\frac{1}{(\sqrt{2\pi\theta_1})^n} e^{\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\theta_1}} \frac{1}{(\sqrt{2\pi\theta_2})^n} e^{\frac{-\sum_{i=1}^{n}(y_i-\mu)^2}{2\theta_2}} \frac{1}{(\sqrt{2\pi\theta_3})^n} e^{\frac{-\sum_{i=1}^{n}(z_i-\mu)^2}{2\theta_3}} \frac{1}{(\sqrt{2\pi\theta_4})^n} e^{\frac{-\sum_{i=1}^{n}(w_i-\mu)^2}{2\theta 4}}$$

$$= (2\pi)^{-2n} (\theta_1\ \theta_2\ \theta_3\ \theta_4)^{-n/2}\ e^{-[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\theta_1}+\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\theta_2}+\frac{\sum_{i=1}^{n}(z_i-\mu)^2}{2\theta_3}+\frac{\sum_{i=1}^{n}(w_i-\mu)^2}{2\theta_4}]}$$

Maximizing the log transformed likelihood function is generally easier.

$$\text{Log } L\ (\boldsymbol{\mu}, \theta_1, \theta_2, \theta_3, \theta_4) = -\ 2n \log (2\pi) - \frac{n}{2} \ln (\theta_1) - \frac{n}{2} \ln (\theta_2) - \frac{n}{2} \ln (\theta_3) - \frac{n}{2} \ln$$

$$(\theta_4) \quad -[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\theta_1}+\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\theta_2}+\frac{\sum_{i=1}^{n}(z_i-\mu)^2}{2\theta_3}+\frac{\sum_{i=1}^{n}(w_i-\mu)^2}{2\theta_4}]$$

Taking the partial derivative with respect to the mean and setting the equation to zero, gives us:

$$\frac{\partial Ln(L)}{\partial\mu} = \frac{2\sum_{i=1}^{n}(xi-\hat{\mu})}{2\theta_1} + \frac{2\sum_{i=1}^{n}(yi-\hat{\mu})}{2\theta_2} + \frac{2\sum_{i=1}^{n}(wi-\hat{\mu})}{2\theta_3} + \frac{2\sum_{i=1}^{n}(hi-\hat{\mu})}{2\theta_4} \underset{=}{\underline{set}} 0$$

The maximum likelihood estimator (MLE) of μ is:

$$\hat{\mu} = \frac{\hat{\theta}_2\hat{\theta}_3\hat{\theta}_4\sum_{i=1}^{n} x_i + \hat{\theta}_1\hat{\theta}_3\hat{\theta}_4\sum_{i=1}^{n} y_i + \hat{\theta}_1\hat{\theta}_2\hat{\theta}_4\sum_{i=1}^{n} z_i + \hat{\theta}_1\hat{\theta}_2\hat{\theta}_3\sum_{i=1}^{n} w_i}{n(\hat{\theta}_2\hat{\theta}_3\hat{\theta}_4 + \hat{\theta}_1\hat{\theta}_3\hat{\theta}_4 + \hat{\theta}_1\hat{\theta}_2\hat{\theta}_4 + \hat{\theta}_1\hat{\theta}_2\hat{\theta}_3)}$$

By the similar procedure, we get estimators of the other parameters:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}\ ;\ \hat{\theta}_2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu})^2}{n}\ ;\quad \hat{\theta}_3 = \frac{\sum_{i=1}^{n}(z_i - \hat{\mu})^2}{n}\ ;$$

$$\hat{\theta}_4 = \frac{\sum_{i=1}^{n}(w_i - \hat{\mu})^2}{n}$$

3.5 Maximum Likelihood Estimator (MLE) under Ha

Under Ha, we had eight equations and eight unknowns. The likelihood function is: *L(μ1, μ2, μ3, μ4, θ1, θ2, θ3, θ4)=* $\prod_{i=1}^{n}$ f(x_i| **μ₁**,θ₁)$\prod_{i=1}^{n}$ f(y_i| **μ₂**,θ₂)$\prod_{i=1}^{n}$ f(z_i| **μ₃**,θ₃)$\prod_{i=1}^{n}$ f(w_i| **μ₄**,θ₄) =

$$\frac{1}{(\sqrt{2\pi\theta_1})^n} e^{\frac{-\sum_{i=1}^{n}(x_i-\mu_1)^2}{2\theta_1}} \frac{1}{(\sqrt{2\pi\theta_2})^n} e^{\frac{-\sum_{i=1}^{n}(y_i-\mu_2)^2}{2\theta_2}} \frac{1}{(\sqrt{2\pi\theta_3})^n} e^{\frac{-\sum_{i=1}^{n}(z_i-\mu_3)^2}{2\theta_3}} \frac{1}{(\sqrt{2\pi\theta_4})^n} e^{\frac{-\sum_{i=1}^{n}(w_i-\mu_4)^2}{2\theta 4}}$$

Maximization of the log transformed likelihood function.

Log L ($\mu_1, \mu_2, \mu_3, \mu_4, \theta_1, \theta_2, \theta_3, \theta_4$) = -2n log (2π) - $\frac{n}{2}$ ln(θ_1)- $\frac{n}{2}$ ln(θ_2) - $\frac{n}{2}$ ln(θ_3) - $\frac{n}{2}$ ln(θ_4)

$$-[\frac{\sum_{i=1}^{n}(x_i - \mu_1)^2}{2\theta_1} + \frac{\sum_{i=1}^{n}(y_i - \mu_2)^2}{2\theta_2} + \frac{\sum_{i=1}^{n}(z_i - \mu_3)^2}{2\theta_3} + \frac{\sum_{i=1}^{n}(w_i - \mu_4)^2}{2\theta_4}]$$

Taking partial derivatives with respect to individual parameters and setting the equations equal to zero gives:

$$\frac{\partial Ln(L)}{\partial \mu_1} = \frac{2\sum_{i=1}^{n}(x_i - \hat{\mu}_1)}{2\theta_1} \overset{set}{=\!=} 0$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} x_i}{n};$$

Similarly, I derived other ML estimators as follows:

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n} y_i}{n} \;;\; \hat{\mu}_3 = \frac{\sum_{i=1}^{n} z_i}{n} \;;\; \hat{\mu}_4 = \frac{\sum_{i=1}^{n} w_i}{n}$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_1)^2}{n};\; \hat{\theta}_2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_2)^2}{n} \;;\; \hat{\theta}_3 = \frac{\sum_{i=1}^{n}(z_i - \hat{\mu}_3)^2}{n};$$

$$\hat{\theta}_4 = \frac{\sum_{i=1}^{n}(w_i - \hat{\mu}_4)^2}{n}$$

3.6 Likelihood ratio test statistic

The ratio of likelihood under Ho true and Ha true is denoted by lambda.

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{\prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{f(xi} \mid \hat{\mu},\ {}_{\mathrm{o}}\hat{\theta}_1) \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{f(yi} \mid \hat{\mu},\ {}_{\mathrm{o}}\hat{\theta}_2) \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{f(zi} \mid \hat{\mu},\ {}_{\mathrm{o}}\hat{\theta}_3) \prod_{\mathrm{i=}}^{\mathrm{n}} \mathrm{f(wi} \mid \hat{\mu},\ {}_{\mathrm{o}}\hat{\theta}_4)}{\prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{f(xi} \mid \hat{\mu}_1, \mathrm{a}\hat{\theta}_1) \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{f(yi} \mid \hat{\mu}_2, \mathrm{a}\hat{\theta}_2) \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{f(zi} \mid \hat{\mu}_3, \mathrm{a}\hat{\theta}_3) \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{f(wi} \mid \hat{\mu}_4, \mathrm{a}\hat{\theta}_4)}$$

Where ${}_a\hat{\theta}_i$, i=1, 2, 3, 4 was variance estimator when Ha was true, and ${}_0\hat{\theta}_i$, i=1, 2, 3, 4 was variance estimator when Ho was true.

$$= \frac{\frac{1}{(\sqrt{2\pi_0\hat{\theta}_1})^n} e^{\frac{-\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{2_0\hat{\theta}_1}} \frac{1}{(\sqrt{2\pi_0\hat{\theta}_2})^n} e^{\frac{-\sum_{i=1}^{n}(y_i-\hat{\mu})^2}{2_0\hat{\theta}_2}} \frac{1}{(\sqrt{2\pi_0\hat{\theta}_3})^n} e^{\frac{-\sum_{i=1}^{n}(z_i-\hat{\mu})^2}{2_0\hat{\theta}_3}} \frac{1}{(\sqrt{2\pi_0\hat{\theta}_4})^n} e^{\frac{-\sum_{i=1}^{n}(w_i-\hat{\mu})^2}{2_0\hat{\theta}4}}}{\frac{1}{(\sqrt{2\pi_a\hat{\theta}_1})^n} e^{\frac{-\sum_{i=1}^{n}(x_i-\hat{\mu}_1)^2}{2_a\hat{\theta}_1}} \frac{1}{(\sqrt{2\pi_a\hat{\theta}_2})^n} e^{\frac{-\sum_{i=1}^{n}(y_i-\hat{\mu}_2)^2}{2_a\hat{\theta}_2}} \frac{1}{(\sqrt{2\pi_a\hat{\theta}_3})^n} e^{\frac{-\sum_{i=1}^{n}(z_i-\hat{\mu}_3)^2}{2_a\hat{\theta}_3}} \frac{1}{(\sqrt{2\pi_a\hat{\theta}_4})^n} e^{\frac{-\sum_{i=1}^{n}(w_i-\hat{\mu}_4)^2}{2_a\hat{\theta}4}}}$$

$$= [\frac{{}_a\hat{\theta}_1 \, {}_a\hat{\theta}_2 \, {}_a\hat{\theta}_3 \, {}_a\hat{\theta}_4}{{}_o\hat{\theta}_1 \, {}_o\hat{\theta}_2 \, {}_o\hat{\theta}_3 \, {}_o\hat{\theta}_4}]^{\frac{n}{2}}$$

(Note: Some step by step algebraic computations are omitted in this derivation).

4. Results

Table 2 contains the empirical alphas for the *F* test and the -2 log (λ). It also contains the robust test (P05) values for the sampling distribution of the likelihood ratio statistic. This table summarizes the rejection rates (the alpha values) for 20,000 trials generated from a certain sample size and variance combination. Note that the P05 values are stable over a spectrum of variance patterns.

The results from table 2 are summarized in figures 1, 2 and 3 below. Figure 1 contains the empirical alpha values for the *F* test as a function of sample size and variance pattern. Figure 2 contains the empirical alpha values for the -2 log (λ) results (as a function of sample size and variance pattern) and figure 3 contains the P05 for the likelihood ratio test statistic, as a function of sample size and variance patterns.

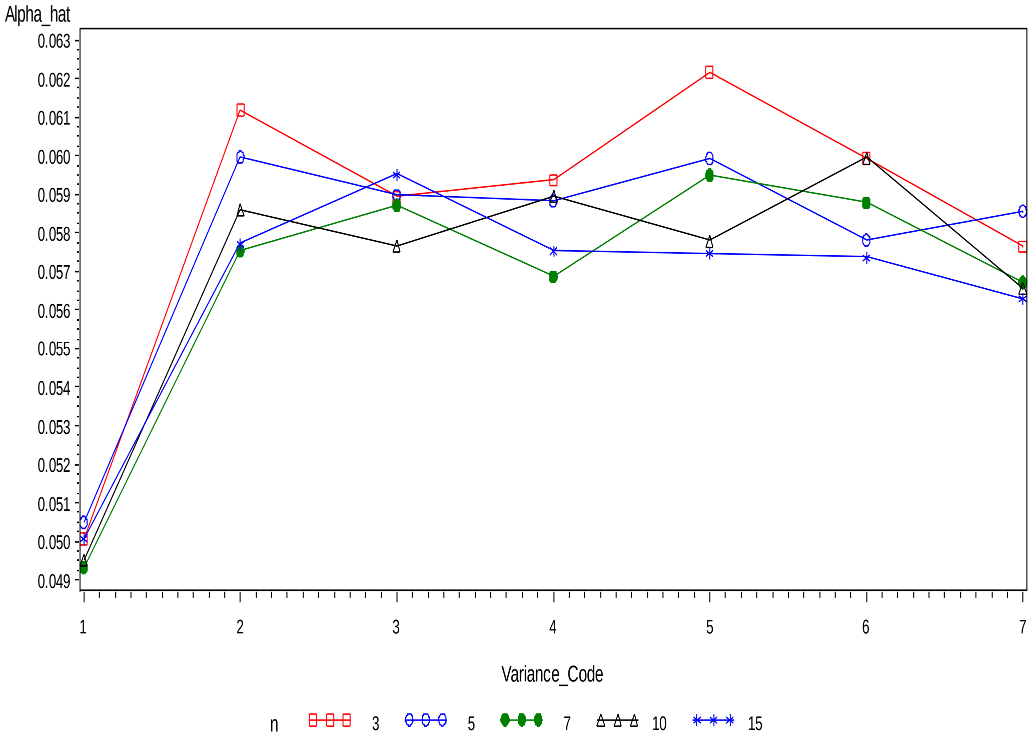

Figure 1: Empirical alpha, *F* test, 20,000 trials per simulation, three simulations per sample size and variance combination

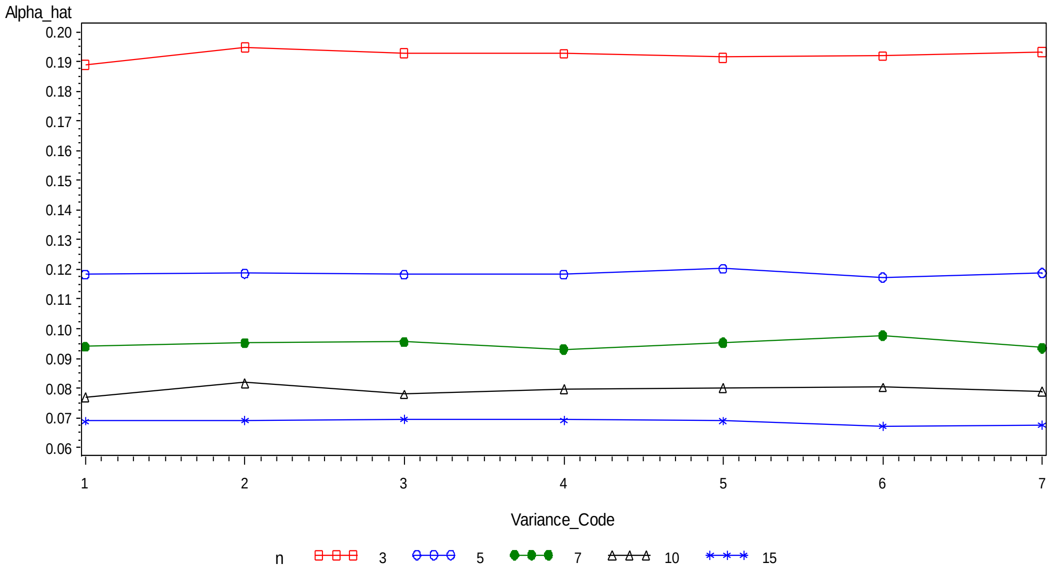

Figure 2: Empirical Alpha, -2 log (λ), 20,000 trials per simulation, three simulations per sample size and variance combination

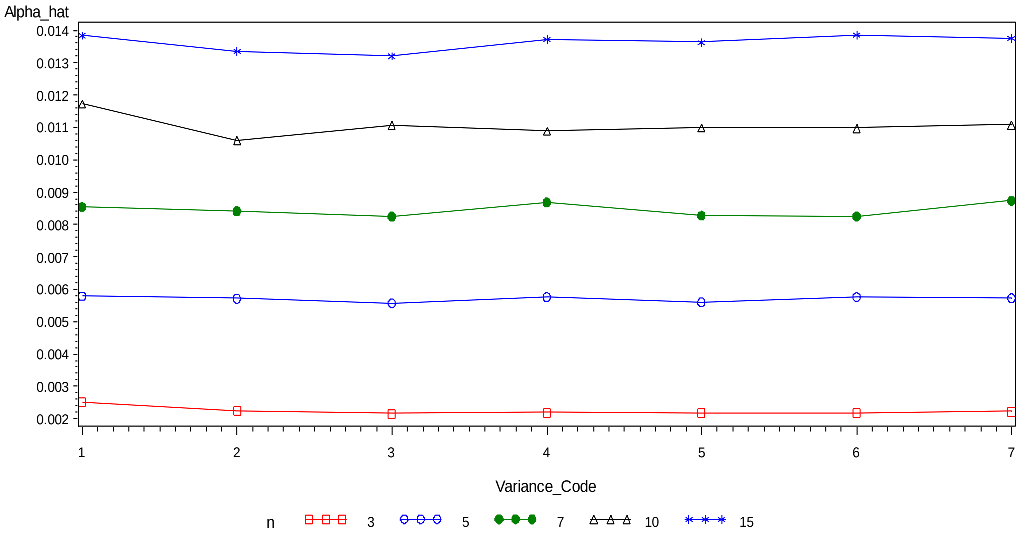

Figure 3: Empirical Alpha, 5 th Percentile, PO5, 20,000 trials per simulation, three simulations per sample size and variance combination

Table 2: Empirical Alpha (Simulated rejection rates) for F test, Chi-Square test, and PO5 Mean value of 3 replications of 20,000 trials (n = 3, 5, 7, 10)

Variance	*n*	Alpha, *F*	Alpha, -2 log(λ)	Alpha, PO5
1,1,1,1	3	0.050083	0.188883	0.002511
1,2,3,4	3	0.061183	0.194743	0.00225
0.25,0.50,0.75,1.00	3	0.058933	0.1929	0.00216
0.5,1.0,1.5,2.0	3	0.059367	0.192667	0.002195
2,4,6,8	3	0.062155	0.191433	0.00218
2.5,5.0,7.5,10.0	3	0.059933	0.191917	0.002172
4,8,12,15	3	0.05765	0.193283	0.002225
1,1,1,1	5	0.0505	0.118283	0.00579
1,2,3,4	5	0.059967	0.118633	0.005714
0.25,0.50,0.75,1.00	5	0.058983	0.118267	0.005571
0.5,1.0,1.5,2.0	5	0.058833	0.1183	0.005763
2,4,6,8	5	0.059933	0.120233	0.005601
2.5,5.0,7.5,10.0	5	0.057817	0.117333	0.005762
4,8,12,15	5	0.058567	0.118767	0.005729
1,1,1,1	7	0.049333	0.093983	0.008546
1,2,3,4	7	0.057533	0.095217	0.008414
0.25,0.50,0.75,1.00	7	0.058717	0.09555	0.008251
0.5,1.0,1.5,2.0	7	0.056867	0.0931	0.008685
2,4,6,8	7	0.0595	0.0953	0.008277
2.5,5.0,7.5,10.0	7	0.058783	0.097767	0.008257
4,8,12,15	7	0.056717	0.09355	0.008733
1,1,1,1	10	0.049533	0.076967	0.011725
1,2,3,4	10	0.0586	0.081817	0.010601
0.25,0.50,0.75,1.00	10	0.057667	0.077983	0.01107
0.5,1.0,1.5,2.0	10	0.058967	0.079833	0.010886
2,4,6,8	10	0.0578	0.08016	0.010989
2.5,5.0,7.5,10.0	10	0.05995	0.080333	0.010981
4,8,12,15	10	0.056567	0.079017	0.011082

Table 2 (Contd.): Empirical Alpha (Simulated rejection rates) for F test, Chi - Square test, and PO5 Mean value of 3 replications of 20,000 trials (n = 15)

Variance	*n*	Alpha, *F*	Alpha, -2 log(λ)
1,1,1,1	15	0.050083	0.068933
1,2,3,4	15	0.057717	0.069217
0.25,0.5,0.75,1	15	0.059517	0.069517
0.5,1,1.5,2	15	0.057533	0.069283
2,4,6,8	15	0.057467	0.069
2.5,5,7.5,10	15	0.057367	0.067233
4,8,12,15	15	0.0563	0.06755

5. Discussions

From my results, we can see that the simulated alpha for the *F* test is close to 0.05 for the equal variance case, but increases above α=0.05 (see figure 1) as the variances for all four populations were different (I have six such variance patterns). The estimated alpha for the six (unequal variance and sample size) combinations hovers around 0.056 – 0.062. Sample sizes essentially have no impact on the simulated alphas *in this current work.*

The simulated alpha for the - 2 log (λ) test is far above 0.05 for any sample size and variance pattern we have studied, but as the sample size increases, the simulated alphas are getting closer to 0.05 (no matter if the variances are equal or unequal) (see figure 2). The different variance patterns have no impact on the simulated alphas within the fixed sample size. In another words, for *n* = j, the simulated alpha remains constant for any variance combination.

Based on my research for a fixed value of *n*, the 5th percentile of lambda (PO5) is relatively stable over differing variance patterns (figure 3). It remains to be seen whether the critical values (P05) can serve as robust rejection regions for data originating from a variety of variance patterns.

This work is a continuation and extension of the work done by Yan, X. (2009). In that study, the simulated alpha for the *F* test was close to 0.05 for the equal variance case, but increased above α=0.05 as the variance increased for population four. The simulated alpha for the -2 log (λ) test was far above 0.05 for small sample sizes, but as the sample size increases, the simulated alphas were somewhat closer to 0.05 (no matter if the variances were equal or unequal).

It should be noted that a mistake was found in the maximization algorithm used by Yan which lowered the original P05 values for *n* = 3 and *n* = 5. A corrected table of P05 values is included below (table 3). For example, the corrected P05 values for *n* = 3 hover around 0.0023 instead of the original 0.001. Also the corrected values for the *n* = 5 case are larger than the originally reported values.

Table 3: Corrected PO5 values (mean of 3 replications of 20,000 trials)

Variance	Simulated Alpha, PO5	
	n =3	*n* =5
1,1,1,1	0.00241289	0.00582421
1,1,1,2	0.00244107	0.00605153
1,1,1,4	0.00229622	0.00603319
1,1,1,6	0.00217265	0.00602717
1,1,1,8	0.00219454	0.00548039
1,1,1,15	0.00214943	0.00579775
1,1,1,30	0.00209895	0.00581112

It is worth comparing my P05 values to Yan's values (n=3 or n = 5 for example) with the caveat that the two sets of generating variances were different in nature. Another comment arises from the visual comparison of the F_{alpha} values of Yan versus my current study. The pattern of three equal variances (with one increasing variance) seems to have a more detrimental effect on the F test than the variance patterns we used. One explanation would be that the F tests (from Yan) were based on a pooled variance estimate from 3 'equal' variances, and hence the larger variance contributed less to the pooled estimate (MSE). As a result Yan's MSEs was generally based on the smaller variance value, and hence she had more power in those F tests.

6. Limitations

My results focus on the case of four normal populations with equal means. I have examined seven different variance patterns and five different sample sizes. I have not studied other variance patterns and more levels of treatment. I also studied only the balanced cases.

7. Future Research

Future research requires that other variance patterns can be evaluated to determine if the new test is robust to more general variance patterns. In addition, a larger number of populations can be evaluated (larger than the four studied currently) and critical values for other alpha levels could be obtained. Further research could focus on some other design structures like incomplete block design, split plot design and some other treatment structures like factorial designs, empty cells (cell means model), etc.

References

Agresti, A. and Finlay, B. (1997). *Statistical Methods for the Social Sciences* (4 th edition). New Jersey: Prentice Hall.

Allison, D., Neale, M., Zannolli, R., Schork, N., Amos, C. and Blangero, J. (1999). Testing the Robustness of the Likelihood-Ratio Test in a Variance-Component Quantitative-Trait Loci-Mapping procedure. *The American Society of Human Genetics*, 65, 531-544.

Cox, T. and Jaber, K. (1990). The generalized likelihood ratio test for the Behrens-Fisher problem. *Journal of Applied Statistics*, 17.1, 63-71.

Hogg, R. and Craig, A. (1995). *Introduction to Mathematical Statistics* (5 th edition). New Jersey: Prentice Hall.

Kuehl, R. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis* (2 nd edition). California: Duxbury.

Rencher, A. and Schaalje, G. (2008). *Linear Models in Statistics* (2 nd edition). N.J.: John Wiley and Sons, Inc.

SAS Institute Inc. (2004). *SAS / STAT ® 9.1 User's Guide*. Cary, N.C., USA: SAS Institute Inc.

Yan, X. (2009). A Robust Likelihood Ratio Test: Testing Equal Means In The Presence of Unequal Variance. *Technical report, College of Economics and International Business, Experimental Statistics Program, New Mexico State University.*