AN INVESTIGATION OF THE POWER OF VARIOUS ALTERNATIVES TO THE ANOVA F STATISTIC WHEN POPULATION VARIANCES ARE UNEQUAL

by

Mark Allan Sorell

B. S., Kansas State University, 1986

A MASTERS REPORT

Submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1988

Approved by:

James J. Higgins

Major Professor

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Dr. James Higgins. His keen guidance, knowledge and patience have taught me far more than any course ever could have.

I am grateful to my committee members, Dr. Mark McNulty and Dr. George Milliken for their critical review of this manuscript.

I would also like to thank Dr. John Boyer and Dr. George Milliken for finding the time to help me get through the difficult times and enjoy the good times a little more.

I would like to thank Dr. Holly Fryer for always having the time to listen, for founding the Department of Statistics at Kansas State University and for making the Fryer Scholarship available.

I wish to express my appreciation to the entire Department of Statistics who have all touched my life in some way during the past five years and given me friends and memories I will treasure for the rest of my life.

And finally I would like to thank my parents, family and friends for their love and support in helping me achieve my goals at Kansas State University.

TABLE OF CONTENTS

	Page
LIST OF TABLES	iii
INTRODUCTION	1
ALTERNATIVE METHODOLOGIES	3
The Analysis of Variance Method	4
The Method of Box	5
The Method of Brown & Forsythe	7
The Method of Welch	9
The Method of James	11
Other Methods	12
REVIEW OF LITERATURE	16
M. B. Brown and A. B. Forsythe (1974)	16
R. L. Kohr and P. A. Games (1974)	21
J. B. Dijkstra and P. S. P. J. Werter (1981)	30
R. R. Wilcox, V. L. Charlin and K. L. Thompson (1986)	35
A. J. Tomarken and R. C. Serlin (1986)	40
CONCLUSIONS AND RECOMMENDATIONS	48
LITERATURE CITED	52
ABSTRACT	54

LIST OF TABLES

		Page
TABLE 1.	Data for Examples	3
TABLE 2.	Empirical Type I Error Probabilities, Nominal size = .05	19
TABLE 3.	Empirical Power of the Tests, Nominal size = .05	20
TABLE 4.	Empirical Type I Error Probabilities, Nominal size = .05	27
TABLE 5.	Estimated Empirical Power of the Tests, Nominal size = .05	28
TABLE 6.	Estimated Empirical Power of the Tests, Nominal size = .05	28
TABLE 7.	Variance Conditions of the Study	28
TABLE 8.	Estimated Empirical Power of the Tests, Nominal size = .05	29
TABLE 9.	Estimated Empirical Power of the Tests, Nominal size = .05	29
TABLE 10.	Empirical Type I Error Probabilities, Nominal size = .05	33
TABLE 11.	Empirical Power of the Tests, Nominal size = .05	34
TABLE 12.	Empirical Type I Error Probabilities, Nominal size = .05	38
TABLE 13.	Empirical Power of the Tests, Nominal size = .05	39
TABLE 14.	Design of the Monte Carlo Investigation	44
TABLE 15.	Empirical Type I Error Probabilities, Nominal size = .05	44
TABLE 16.	Empirical Power of the Tests, Nominal size = .05	45
TABLE 17.	Empirical Power of the Tests, Nominal size = .05	46-47

INTRODUCTION

Testing the equality of two or more population means is a problem that occurs in nearly all disciplines. The usual analysis of variance F test for equality of means is based on three assumptions:

(1) independent random samples are selected from the populations,

(2) the data from each population are normally distributed, and

(3) the population variances are equal.

This study focus's on the violations of assumption (3), the equal variance assumption. Various alternatives to the usual analysis of variance F test which have appeared in the literature will be presented and their properties discussed. Recommendations will be made concerning the use of these tests.

For an experimenter trying to choose an acceptable alternative for the usual analysis of variance F test there are two topics of concern, robustness and power. Robustness refers to the ability of the test statistic to hold the Type I error rate at the desired level when basic assumptions are violated. The power of a test statistic is its ability to detect differences among population means when there are differences among them.

If a test is to be applied in a range of circumstances in which assumptions are violated, the test must be robust. Lack of control of Type I error rates will make a test unacceptable to the applied researcher. Among tests that are robust, power can be used as a criterion for choosing among them.

One of the common beliefs about the unequal variances situation is that if the treatments have equal sample sizes then the usual ANOVA F test is satisfactory in some sense. The findings of this report bring the conventional wisdom into question. A most important point for a researcher is not to choose an alternative test solely on the basis of robustness but under careful consideration of the power of the test statistic also. The primary focus of this report is the power of alternatives to the usual ANOVA F test when the equality of variance assumption is violated.

ALTERNATIVE METHODOLOGIES

Let x_{ij} be the j^{th} observation in the i^{th} group, where $j = 1,\ldots,n_i$ and $i = 1,\ldots,k$. The x_{ij}'s are assumed to be independent and normally distributed with expected values μ_i and variances σ_i^2. The minimum variance unbiased estimates of μ_i and σ_i^2 are,

$\bar{x}_{i.} = \sum_{j=1}^{n_i} x_{ij}/n_i$ and $S_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})/(n_i - 1)$ respectively.

Various alternatives to the usual ANOVA F test are presented. A numerical example is given to illustrate the computation involved for each test. In order to demonstrate numerically how to perform the various tests the data set in Table 1 was used in each case.

TABLE 1. Data for Examples.

Group	1	2	3	4
	1	12	12	13
	8	10	4	14
	9	13	11	14
	9	13	7	17
	4	12	8	11
	0	10	10	14
	1		12	13
			5	14
n_i	7	6	8	8
$\sum_{j=1}^{n_i} x_{ij}$	32	70	69	110
$\bar{x}_{i.}$	4.571	11.667	8.625	13.750
S_i^2	16.286	1.867	9.696	2.786

1) THE ANALYSIS OF VARIANCE METHOD

The usual analysis of variance F test for testing the equality of population means in a completely randomized design, with one-way classification, is given by the following equation,

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2/(k-1)}{\sum_{i=1}^{k} (n_i - 1)\, s_i^2/(N-k)}$$

where $N = \sum_{i=1}^{k} n_i$ and $\bar{x}_{..} = \sum_{i=1}^{k} \sum^{i}_{j=1} x_{ij}/N$.

When all population means and variances are equal, this statistic follows the F distribution with (k - 1) and (N - k) degrees of freedom respectively. This will be referred to as the ANOVA F statistic.

An Example of ANOVA

An example of computations for the ANOVA F statistic are calculated using the data in TABLE 1,

$$N = 7+6+8+8 = 29 \qquad \bar{x}_{..} = 281/29 = 9.690$$

$$F = \frac{[7(4.571-9.69)^2+6(11.667-9.69)^2+8(8.625-9.69)^2+8(13.75-9.69)^2]/3}{[6(16.286) + 5(1.867) + 7(9.696) + 7(2.786)]/25}$$

$$= 115.941/7.777 = 14.908.$$

The ANOVA F value of 14.908 would be compared to a critical value denoted as $F_{\alpha}(3,25)$ where α is the probability of a Type I error chosen before hand, and 3 and 25 are the degrees of freedom of the numerator and denominator, respectively. Thus $F_{.05}(3,25) = 2.991$, so

the null hypothesis that all means are equal would be rejected at the .05 level of significance.

2) THE METHOD OF BOX

Box (1954) proposed a procedure that requires the computation of the ANOVA F Statistic but adjusts the critical value and the degrees of freedom to account for the unequal variances. Box proved that a bias coefficient, b, determines the direction of the discrepancy between the actual probability of the Type I error rate and the nominal Type I error rate. The b coefficient is approximately the ratio of the unweighted and weighted means of the population variances. When the group sample sizes (n_i's) are equal, the weighted and unweighted mean variances are equal, hence $b = 1.0$. When the group sample sizes are unequal, unless the variances are homogeneous, b may be either greater than or less than 1.0. When $b \neq 1.0$, the ANOVA F statistic will be biased. Box showed that the mean square ratio of the ANOVA F statistic is approximately distributed as $bF(h',h)$ where h' and h represent reduced degrees of freedom and F represents an F random variable. Although Box defines b, h' and h from the population variances it is possible to substitute the estimated variances from the sample data in the following equations:

$$\tilde{b} = \frac{(N-k)}{N(k-1)} \left[\sum_{i=1}^{k} (N - n_i)s_i^2 \right] / \left[\sum_{i=1}^{k} (n_i - 1)s_i^2 \right]$$

$$h' = \left[\sum_{i=1}^{k} (N - n_i)s_i^2 \right]^2 / \left[\left(\sum_{i=1}^{k} n_i s_i^2 \right)^2 + N \sum_{i=1}^{k} (N - 2n_i)s_i^4 \right]$$

$$h = \left[\sum_{i=1}^{k} (n_i - 1)s_i^2 \right]^2 / \left[\sum_{i=1}^{k} (n_i - 1)s_i^4 \right].$$

An Example of Box's Method

An example of computations for Box's method are calculated using the data in TABLE 1,

$$\tilde{b} = \frac{(25)}{29(3)} \frac{[22(16.286) + 23(1.867) + 21(9.696) + 21(2.786)]}{[\ 6(16.286) + \ 5(1.867) + \ 7(9.696) + \ 7(2.786)]}$$

$$= [25/29(3)][663.355/194.425] = 0.980$$

$$h' = \frac{[22(16.286) + 23(1.867) + 21(9.696) + 21(2.786)]^2}{\{[7(16.286) + \ 6(1.867) + \ 8(9.696) + \ 8(2.786)]^2 + 29[15(16.286)^2 + 17(1.867)^2 + 13(9.696)^2 + 13(2.786^2]\}}$$

$$= (663.355)^2/[(225.06)^2 + (29)5360.828] = 2.135$$

$$h = \frac{[6(16.286) + 5(1.867) + 7(9.696) + 7(2.786)]^2}{[6(16.286)^2 + 5(1.867)^2 + 7(9.696)^2 + 7(2.786)^2]}$$

$$= (194.425)^2/2321.251 = 16.285$$

$$bF_{.05}(h',h) = bF_{.05}(2.135,16.285) = .980(3.552) = 3.481.$$

Recall the ANOVA F = 14.908. Box's method yields a critical value of 3.481 so the null hypothesis would be rejected. Most computer packages will do decimal degrees of freedom. A conservative critical value could be obtained by checking the critical values corresponding to integer degrees of freedom on either side of the fractional degrees

of freedom and chosing the largest value. For this example 2 and 16 degrees of freedom would be the conservative degrees of freedom.

It should be noted that in the equal sample size case, after some algebraic manipulation, the critical value reduces to $F_{\alpha}\{(k - 1)\epsilon', (N - k)\epsilon\}$ where ϵ' and ϵ, the factors by which the degrees of freedom are reduced, are given by

$$\epsilon' = 1/\{1 + ((k - 2)/(k - 1))c^2\} \quad , \quad \epsilon = 1/(1 + c^2)$$

and c is the coefficient of variation of the variances. That is to say,

$$c^2 = (1/k) \sum_{i=1}^{k} (\sigma_i^2 - \bar{\sigma}^2)^2/(\bar{\sigma}^2)^2 \ ,$$

where

$$\bar{\sigma}^2 = \sum_{i=1}^{k} \sigma_i^2/k \ .$$

Because the population variances (σ_i^2's) are unknown, use the estimated variances (S_i^2's) to estimate the parameters.

3) THE METHOD OF BROWN & FORSYTHE

Brown & Forsythe (1974) suggest a statistic in which the numerator is the same as the ANOVA F statistic. The difference is in the denominator which has expectation equal to the numerator when all means are equal. That is,

$$F^* = \frac{\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2}{\sum_{i=1}^{k} (1 - n_i/N)s_i^2} .$$

Critical values are obtained from the F distribution with (k - 1) and f degrees of freedom, respectively, where

$$1/f = \sum_{i=1}^{k} c_i^2/(n_i - 1) \quad \text{and} \quad c_i = \frac{(1 - n_i/N)s_i^2}{\sum_{i=1}^{k} (1 - n_i/N)s_i^2} .$$

Brown & Forsythe used the Satterthwaite (1941) approximation for f. When there are only two groups Brown & Forsythe's statistic reduces to what is known as the Welch (1936) approximate degrees of freedom solution to the Behrens-Fisher problem. Although Scheffe'(1944) proved that exact solutions of this type cannot be found, a simulation study by Wang (1971) has shown this approach is adequate for the size of the test for the k = 2 case.

An Example of Brown & Forsythe's Method

An example of computations for Brown and Forsythe's method are calculated using the data in TABLE 1,

$$F^* = \frac{[\,7(4.571-9.69)^2 + 6(11.667-9.69)^2 + 8(8.625-9.69)^2 + 8(13.75-9.69)^2\,]}{[\,(1-7/29)16.286 + (1-6/29)1.867 + (1-8/29)9.696 + (1-8/29)2.786\,]}$$

$$= 347.823/22.874 = 15.206$$

$$c_1 = (1 - 7/29)16.286/22.874 = 0.540$$

$$c_2 = (1 - 6/29)\ 1.867/22.874 = 0.065$$

$$c_3 = (1 - 8/29)\ 9.696/22.874 = 0.307$$

$c_4 = (1 - 8/29)\ 2.786/22.874 = 0.088$

$1/f = [.540^2/6 + .065^2/5 + .307^2/7 + .088^2/7] = 0.064$, $f = 15.625$.

Since Brown & Forsythe's method yields $F^* = 15.206$, compared to $F_{.05}(3, 15.625) = 3.256$ the null hypothesis would be rejected.

4) THE METHOD OF WELCH

Welch (1951) suggested the following statistic,

$$v^2 = \frac{\sum_{i=1}^{k} w_i(\bar{x}_{i.} - \tilde{x}_{..})^2/(k - 1)}{[1 + (2/3)(k - 2)\Lambda]}$$

where

$$w_i = n_i/s_i^2 \quad , \quad \bar{w} = \sum_{i=1}^{k} w_i \quad , \quad \tilde{x}_{..} = \frac{\sum_{i=1}^{k} w_i \bar{x}_{i.}}{\bar{w}} \quad \text{and}$$

$$\Lambda = \frac{3 \sum_{i=1}^{k} (1 - w_i/\bar{w})^2/(n_i - 1)}{(k^2 - 1)} .$$

The numerator of Welch's statistic differs from the ANOVA F numerator in the sense that it weights the overall mean and the deviations from it by w_i rather than n_i. The critical values may be obtained from an F distribution with $(k - 1)$ and $(1/\Lambda)$ degrees of freedom, respectively. When there are only two groups Welch's statistic also reduces to what is known as the Welch approximate degrees of freedom solution to the Behrens-Fisher problem.

An Example of Welch's Method

An example of computations for Welch's method are calculated using the data in TABLE 1,

$w_1 = 7/16.286 = 0.430 \qquad w_2 = 6/1.867 = 3.214$

$w_3 = 8/ 9.696 = 0.825 \qquad w_4 = 8/2.786 = 2.872$

$\bar{w} = (.43 + 3.214 + .825 + 2.872) = 7.341$

$\Lambda = 3\{(1 - .43/7.341)^2/6 + (1 - 3.214/7.341)^2/5 + (1 - .825/7.341)^2/7$

$+ (1 - 2.872/7.341)^2/7\}/15 = 3\{.376\}/15 = 0.075$

$1/\Lambda = 1/.075 = 13.333$

$\tilde{x}_{..} = [.43(4.571) + 3.214(11.667) + .825(8.625) + 2.872(13.75)]/7.341$

$= 86.069 / 7.341 = 11.724$

$v^2 = \{[.43(4.571-11.724)^2 + 3.214(11.667-11.724)^2 + .825(8.625-11.724)^2$

$2.872(13.75-11.724)^2]/3\}/[1 + (2/3)2(.075)]$

$= (41.723/3)/1.100 = 12.644.$

Since Welch's method yields $v^2 = 12.644$, compared to an $F_{.05}(3,13.333) = 3.387$ the null hypothesis would be rejected.

Levy (1978) proposed that the non-null distribution of Welch can be approximated by a non-central F distribution with parameters $(k - 1)$, f'', and λ, where

$$f'' = (k^2 - 1) / 3 \Delta \qquad \Delta = \sum_{i=1}^{k} (1 / (n_i - 1))(1 - w_i/ \bar{w})$$

$$w_i = n_i / \sigma_i^2 \qquad \bar{w} = \sum_{i=1}^{k} w_i$$

$$\bar{\mu}' = \sum_{i=1}^{k} w_i \mu_i / \bar{w} \qquad \text{and} \qquad \lambda = \sum_{i=1}^{k} w_i (\mu_i - \bar{\mu}')^2 .$$

Monte Carlo techniques were used to demonstrate that this approximation is reasonable. Thus, as is the case with an ANOVA, one could determine appropriate sample sizes for achieving a desired level of power associated with Welch's test or, for specific sample sizes, one could determine the power of Welch's test for particular alternatives to the null hypothesis.

5) THE METHOD OF JAMES

James (1951) found a test statistic similar to Welch which differs primarily in its approximations for the critical values. The test statistic proposed by James is simply the numerator of Welch's statistic and may be written as

$$J = \sum_{i=1}^{k} w_i(\bar{x}_{i.} - \tilde{x}_{..})^2/(k - 1)$$

where $w_i = n_i/S_i^2$, $\bar{w} = \sum_{i=1}^{k} w_i$ and $\tilde{x}_{..} = \sum_{i=1}^{k} w_i\bar{x}_{i.}/\bar{w}$.

The critical value is $\chi_\alpha^2 \, h(\alpha)$ where χ_α^2 is the $(1 - \alpha)$ percentile from the chi-square distribution based on $(k - 1)$ degrees of freedom and

$$h(\alpha) = \{1 + [(3\chi_\alpha^2 + (k + 1))/2(k^2 - 1)][\sum_{i=1}^{k} (1 - w_i/\bar{w})^2/(n_i - 1)]\}.$$

Just like Brown & Forsythe's and Welch's statistic, James' statistic also reduces to what is known as the Welch approximate degrees of freedom solution to the Behrens-Fisher problem in the two sample case.

An Example of James' Method

An example of computations for James' method are calculated using the data in TABLE 1,

$w_1 = 7/16.286 = 0.430$ $\quad$ $w_2 = 6/1.867 = 3.214$

$w_3 = 8/ 9.696 = 0.825$ $\quad$ $w_4 = 8/2.786 = 2.872$

$\bar{w} = [.43 + 3.214 + .825 + 2.872] = 7.341$

$\tilde{x}_{..} = [.43(4.571) + 3.214(11.667) + .825(8.625) + 2.872(13.75)]/7.341$

$= 86.069/7.341 = 11.724$

$J = [.43(4.571-11.724)^2 + 3.214(11.667-11.724)^2 + .825(8.625-11.724)^2$

$+ 2.872(13.75-11.724)^2] = 41.723/3 = 13.908.$

The $\alpha = 0.05$ chi-square critical value baseed on 3 degrees of freedom is 7.815,

$h(\alpha) = (1 + [(3(7.815) + 5)/2(15)][(1 - .43/7.341)^2/6$

$+ (1 - 3.214/7.341)^2/5 + (1 - .825/7.341)^2/7 + (1 - 2.872/7.341)^2/7])$

$= [1 + (28.445 / 30)(0.376)] = 1.357$

$\chi^2_{.05}h(.05) = 7.815(1.357) = 10.605.$

Since James' method yields a chi-square value of 13.908, compared to a critical value of 10.605 the null hypothesis would be rejected.

OTHER METHODS

Several techniques exist that were not chosen for a more detailed investigation and comparison either because they were too complicated

to be feasibly used by the typical researcher or little was found on the power behavior of the test statistic.

The Second Order Method of James

One such technique is known as the second order method of James. This method includes a second order approximate term that is added to the correction factor $h(\alpha)$ for the critical value. For $k > 2$ James proposes to use the first order method for smaller samples or the usual chi-square critical value for large samples. In his opinion it would involve too much numerical calculation to include the second order correction term when considering the small gain in precision when the second order term is added into the equation. It should be noted that in 1951 the computers were not as efficient as they are today. Thus, if James' second order correction was implemented in a statistical package so that hand calculation would not have to be done, then the second order method could give slightly better approximations than the first order method.

The Method of Unweighted Means

The method of unweighted means is another technique that has been widely used in recent years. However, Milliken & Johnson (1984) do not recommend its use when the variances are unequal. The test statistic is given by,

$$F = \frac{\tilde{n} \sum_{i=1}^{k} (\bar{x}_{i.} - \hat{x}_{..})^2/(k - 1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2/(N - k)}$$

where, $1/\tilde{n} = (1/k) \sum_{i=1}^{k} (1/n_i)$ and $\hat{x}_{..} = \sum_{i=1}^{k} \bar{x}_{i.}/k$.

The quantity $\tilde{n}$ is the harmonic mean of the sample sizes. The critical values may be obtained from the F distribution with (k - 1) and (N - k) degrees of freedom, respectively. This analysis yields reasonable approximations to the F distribution only when the sample sizes are not too unequal. A theoretical analysis suggests that the size of this technique will be even more affected by heterogeneous variances (when the sample sizes are unequal) than the usual ANOVA F statistic (Kohr & Games 1974). For this reason the method of unweighted means was not considered in the detailed comparisons. It should be noted that when the sample sizes are equal this analysis is identical to ANOVA.

The Method of Two Stage Sampling

Bishop & Dudewicz (1978) present procedures, with tables and approximations needed for implementation, which give exact tests with power and size completely independent of the unknown variances. As a historical note, two-stage sampling procedures were first introduced by Stein (1945) in an equal variance context. The procedure of Bishop & Dudewicz guarantees that the probability of a Type I error is exactly α, and the power is exactly $(1 - \beta)$ for a given value of δ (where $\delta = \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2$) chosen by the researcher.

The primary purpose of the first stage of the procedure is to obtain estimates of the k variances based on n observations randomly

chosen from each treatment group. Once the sample variances are computed, it is possible to determine N_i, the total number of observations needed from the i^{th} group, so that the desired power will be obtained. The second stage consists of sampling the additional $(N_i - n)$ observations that are required for the i^{th} group and then testing the null hypothesis $\mu_1 = \mu_2 = \ldots = \mu_k$.

A practical problem with this method is the requirement of equal sample sizes in the first stage. Work by Wilcox (1987) proposes a simple yet accurate method for handling unequal sample sizes in the first stage of the Bishop & Dudewicz method. If obtaining additional observations is impractical, the procedure by Wilcox might still be useful since it can be used to determine whether the existing sample sizes are reasonably large enough to obtain the desired power. This method was not considered because of its complexity, the requirement of obtaining additional samples, and the lack of literature which would allow comparison to the other procedures under consideration. On the other hand, if the researcher does have the luxury of obtaining additional samples for each treatment then this method is possibly attractive in its ability to control exactly both Type I error rates and power.

REVIEW OF LITERATURE

In this section five papers are reviewed that compare the performance of the test statistics described in the Alternative Methodologies section. The tests are appraised in terms of the Type I error rates and the power under various combinations of sample sizes, variances and alternative hypotheses. The discussion of each paper is divided into four sections: Purpose and Method, Size, Power and Comments.

M. A. Brown and A. B. Forsythe (1974)

Purpose and Method

Brown and Forsythe compared the performance of their test statistic with the test of Welch ,the first order test of James and the ANOVA F statistic. Using four groups, six groups and ten groups the size of each test was studied. The sample sizes ranged from four to twenty-one and the standard deviations ranged from one to three. For the power study, four groups with sample sizes (11,16,16,21) were simulated with equal and unequal variances and four different mean structures. For each set of criterion 10,000 independent replications were simulated.

Size

The ANOVA F statistic shows some considerable deviations from its nominal size when the sample sizes of the groups are unequal. At the 5% level in the examples shown, the empirical size of the ANOVA F varies from 3% when the larger sample sizes are paired with the larger

variances to 17% when the smaller sample sizes are paired with the larger variances. For small sample sizes, the test of James deviates more widely from the nominal size, rejecting the null hypothesis a little too often. Overall the Type I error rate of the Brown-Forsythe test varies slightly more than the test of Welch. For groups with more than ten experimental units the difference between the nominal and empirical sizes of both the Brown-Forsythe and Welch test are small with Welch's test remaining slightly closer to the nominal value, in most cases. Results of the size investigation are shown in TABLE 2.

Power

The power results from this study are given in TABLE 3. The ANOVA F values were given only when the equal variance assumption was met. Because the Welch test and the test of James have similar numerators and Welch's test had better control of the Type I error rate the power calculations from James' test were omitted. When the variances were equal both the Welch test and the Brown-Forsythe test had only slightly less power than the ANOVA F. The Brown-Forsythe test showed higher power, around 10%, only when an extreme mean was paired with the largest variance. In all other cases The Welch test had superior power. For example, when an extreme mean was paired with the smallest variance the gain in power was as high as 35%. When extreme means coincided with the largest and smallest variance as much as a 26% gain in power was obtained by using Welch's test.

Comments

The combination of sample sizes and standard deviations adequately demonstrated the size of the various test statistics in different situations. However, the power study was limited to include only one sample size combination. Other combinations would have been helpful had they been investigated. Also, the power of the ANOVA F would have been interesting to see even though the Type I error rate was not close to the nominal level.

TABLE 2. Empirical Type I Error Probabilities, Nominal size = .05.

Sample Size Condition	Standard Deviation Condition	ANOVA F	Brown-Forsythe	Welch	James
(4,4,4,4)	(1,1,1,1)	.049	.034	.045	.079
	(1,2,2,3)	.067	.041	.047	.081
(4,8,10,12)	(1,1,1,1)	.051	.048	.057	.067
	(1,2,2,3)	.030	.057	.049	.056
	(3,2,2,1)	.144	.062	.065	.077
(11,11,11,11)	(1,1,1,1)	.051	.049	.051	.055
	(1,2,2,3)	.063	.057	.050	.054
(11,16,16,21)	(1,1,1,1)	.049	.051	.050	.053
	(3,2,2,1)	.108	.062	.055	.058
	(1,2,2,3)	.040	.065	.054	.056
(4,4,4,4,4,4)	(1,1,1,1,1,1)	.049	.034	.061	.095
	(1,1,2,2,3,3)	.083	.049	.070	.109
(4,6,6,8,10,12)	(1,1,1,1,1,1)	.046	.045	.061	.074
	(1,1,2,2,3,3)	.031	.065	.062	.075
	(3,3,2,2,1,1)	.170	.058	.068	.084
(6,6,6,6,6,6)	(1,1,2,2,3,3)	.071	.052	.057	.073
(11,11,11,11,11,11)	(1,1,2,2,3,3)	.073	.065	.057	.062
(16,16,16,16,16,16)	(1,1,2,2,3,3)	.072	.068	.051	.052
(21,21,21,21,21,21)	(1,1,2,2,3,3)	.069	.065	.048	.049
(20,20,20,20,20,20 20,20,20,20)	(1,1,1.5,1.5,2, 2,2.5,2.5,3,3)	.071	.066	.052	.053

TABLE 3. Empirical Power of the Tests, Nominal Size = .05, Sample Size Condition for all Cases (11,16,16,21).

Variance Condition	Mean Structure	ANOVA F	Brown-Forsythe	Welch
(1,1,1,1)	(0,0,0,0)	.049	.051	.050
	(1,0,0,0)	.686	.676	.650
	(0,0,0,0.7)	.553	.544	.523
	(0.5,0,0,0.5)	.336	.333	.318
(3,2,2,1)	(0,0,0,0)		.062	.055
	(1.5,0,0,0)		.332	.222
	(0,0,0,1)		.227	.478
	(1.3,0,0,1.3)		.424	.682
(1,2,2,3)	(0,0,0,0)		.065	.054
	(1.3,0,0,0)		.291	.655
	(0,0,0,1)		.273	.175
	(1,0,0,1)		.298	.400

R. L. Kohr and P. A. Games (1974)

Purpose and Method

Kohr and Games study the size and power of the ANOVA F, Welch test and the test by Box. The study of the sizes of the various tests was done with four groups. The sample sizes ranged from six to fourteen and the variances ranged from one to thirteen with nine different sets of variance combinations. The power study included several plots of the power of the tests displayed in graphs at several different levels of the noncentrality parameter. Three different mean structures were investigated with the same variance and sample size conditions that were included in the size study. For each set of criterion, four blocks of 500 replications were used in the simulation.

Size

Ranking the procedures from poor to good TABLE 4 shows the ANOVA F had the worst control of the size. Next was Box's test and finally Welch's test was the best. For the equal sample size case when there was one extreme variance the ANOVA F performed particularly poorly with the size of the test doubling the desired level. Box's test performed similarly with one extreme variance but was not as liberal as the ANOVA F. When the groups had several different variances Box's test was somewhat conservative. Welch's test by far demonstrated the best control of the Type I error rate with empirical probabilities ranging from 5% to 6.3%.

For the unequal sample size cases the ANOVA F tended to be quite conservative when the extreme variance was paired with the largest

sample size and was quite liberal when the extreme variance was paired with the smallest variance. The minimum and maximum empirical values for the ANOVA F were 2.2% and 23.6% respectively. Box's test and Welch's test both performed more adequately than the ANOVA F. Overall the test by Welch was slightly better in the unequal sample size cases because of its smaller range of empirical values 4% to 6.3% as compared to the range of Box's test 3.6% to 7.6%.

Power

Power functions are displayed in TABLES 5, 6, 8, and 9. These values were read from graphs. The power functions were expressed in terms of the noncentrality parameter,

$$\phi = [(n \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2/k) \; / \; \sigma^2 \;]^{1/2}$$

where

$$\bar{\mu} = \sum_{i=1}^{k} \mu_i/k \; .$$

Since all variance conditions used had an average of four and all sample size conditions had a harmonic mean of eight these numbers were used in the above formula.

When the assumption of equal variances was met the ANOVA F statistic was more powerful than either the Welch test or the Box test but the increase in power was small, ranging from 2.5% to 5.5%. In TABLE 6 it is shown that when the means are evenly spaced apart and the sample sizes are equal then Welch's test was the most powerful with Box's test having a severe power loss. The gain in power for the conditions described above was as high as 47%. The power for less

extreme variance conditions are intermediate between those shown in TABLE 6 but the same trends hold whenever the means are evenly spaced apart and the sample sizes are equal. Unfortunately, when the null hypothesis is violated in other ways, the power depends upon what variances accompany the deviant means. As demonstrated previouly by Brown and Forsythe, TABLE 5 shows that when small variances are paired with the more deviant means Welch's test was as high as 34% more powerful than Box's test and as high as 24% more powerful than the ANOVA F statistic. When larger variances are paired with the more deviant means the ANOVA F statistic was the most powerful while Box's test had slightly higher power than Welch's test, as shown in TABLE 5. Caution must be used in this situation because the superiority of the ANOVA F may be inflated due to the fact that the probability of a Type I error may be twice as high as desired because of its lack of control. As TABLE 5 shows, in some situations Box's test does have more power than Welch's test but this gain in power is only as high as 11%, whereas, when reverse conditions hold Welch's test provides gains in power as high as 34%. In none of the equal sample size cases does Box's test have superior power over both the ANOVA F and Welch's test. Thus as Kohr and Games stated, "The only absolute statement that can be made about the equal sample size case when population variances are unknown is that the Box test would not be the preferred test."

When sample sizes are unequal and the variances are unequal, the ANOVA F statistic's failure to control the probability of a Type I error makes its use questionable. When the bias coefficient of Box's

test is less than one ($b < 1.0$), see TABLE 7, and the ANOVA F is conservative, it might still be possible that the ANOVA F would overcome the initial conservative bias and be the most powerful alternative for medium to large values of the noncentrality parameter. If the experimenter is willing to accept a 5% risk of a Type I error, they should have little complaint if the actual size is 1% and the test is still the most powerful of any available. It is only when ($b > 1.0$) and the ANOVA F is biased in the liberal direction that the ANOVA F must be discarded.

When the larger variances were paired with the larger sample sizes ($b < 1.0$), if the alternative hypothesis had means that were equally spaced apart or deviant means paired with smaller variances and smaller sample sizes, then Welch's test was consistently the most powerful of the three tests. Within each alternative hypothesis the superiority of the Welch test roughly decreased as the bias coefficient (b) approached 1.0.

Under the alternative where ($\mu_3 < \mu_1 = \mu_2 < \mu_4$) the results were mixed and not as clear. When the variances were (1,1,1,13) the results were similar to those above with Welch's test being the most powerful, but when the variances were (1,2,3,10) or (1,3,5,7) the Box test was the most powerful by as much as 18% as shown in TABLE 8. To be assured of using the most powerful test the experimenter must know whether ($b > 1$) or ($b < 1$) and the kind of alternative hypothesis that is expected. But this is a most unlikely situation for the use of an omnibus test like the three discussed in this paper. If in fact the experimenter anticipates on an a priori basis that the alternative

hypothesis would be ($\mu_3 < \mu_1 = \mu_2 < \mu_4$) usually they would perform appropriate contrasts to gain greater power rather than using an omnibus test.

When larger population variances are paired with the smaller sample sizes ($b > 1.0$) results as shown in TABLE 9 apply. The complex results for this variance and sample size condition occured when the alternative hypothesis was ($\mu_1 < \mu_3 = \mu_4 < \mu_2$). For the equally spaced means alternative and when the extreme means were paired with the larger sample sizes and smaller variances the Welch test was consistently more powerful than the Box test. However, for the ($\mu_1 < \mu_3 = \mu_4 < \mu_2$) alternative, the Box test was more powerful than Welch's test for the variance condition of (7,5,3,1) but less powerful for the (13,1,1,1) variance condition as shown in TABLE 9.

Overall the Welch test demonstrated superb control of the Type I error rate and usually had power superior to Box's test. The only times that the Box test demonstrated greater power were on the few occasions when two means deviated largely from the grand mean and both of the means were paired with relatively large variances. As noted by Kohr and Games, many unequal variance conditions produced results where the power superiority of the Welch test was even greater than the 47% shown in TABLE 9, while the power superiority of the Box test over Welch's test never exceeded more than 13%.

Kohr and Games suggest that one could make use of Box's bias coefficient (b) as follows. If the bias coefficient is between 0.88 and 1.05 then the ANOVA F would be used and if the value was outside

this range then the Welch test would be used. The authors also suggested another way to use Box's test in an omnibus procedure, but felt that such a procedure would be too complex for the everyday experimenter. The authors conclude that the Welch test is the test of choice when the sample sizes are unequal, but it is not better than the ANOVA F in the equal sample size case.

Comments

This paper by Kohr and Games was the only paper found that demonstrated the performance of both Box's test and Welch's test under numerous sample size and variance conditions. The size study was conducted with nine different variance conditions and showed just how poorly the ANOVA F controls the size when the variances are (1,1,1,13) or (13,1,1,1) and the sample sizes are equal. There is some doubt about the author's reccommendation to use the ANOVA F when the sample sizes are equal. Although the ANOVA F did have superior power, its inability to control the Type I error rate makes its use questionable. With the above variance condition the empirical Type I error rate could be twice as high as desired and some researchers may find this objectionable. The increase in power may not be worth the risk that is involved. The findings of this paper about the performance of Welch's test are consistent with the previous paper. New findings included the comparisons of Box's test with Welch's test. For an omnibus test to be run on any particular set of data where the assumption of equal variances is violated the Welch test should be performed because of its excellent control of the size. There is little doubt that in some situations this test will not be the most

powerful but on these rare occasions the power loss would be from 1% to 25%.

TABLE 4. Empirical Type I Error Probabilities, Nominal Size = .05.

Sample Size Condition	Variance Condition	ANOVA F	Box	Welch
(8,8,8,8)	(1,1,1,13)	.115	.070	.063
	(1,2,3,10)	.074	.050	.050
	(1,3,5,7)	.065	.041	.056
	(2.67,4,4,5.33)	.051	.044	.050
	(4,4,4,4)	.052	.045	.048
	(5.33,4,4,2.67)	.063	.053	.059
	(7,5,3,1)	.060	.045	.054
	(10,3,2,1)	.073	.044	.053
	(13,1,1,1)	.099	.055	.053
(6,8,9,9)	(1,1,1,13)	.074	.048	.050
	(1,2,3,10)	.056	.052	.057
	(1,3,5,7)	.057	.050	.054
	(2.67,4,4,5.33)	.049	.045	.057
	(4,4,4,4)	.052	.046	.055
	(5.33,4,4,2.67)	.058	.045	.050
	(7,5,3,1)	.081	.049	.059
	(10,3,2,1)	.114	.067	.060
	(13,1,1,1)	.160	.067	.053
(5,7,10,14)	(1,1,1,13)	.030	.053	.055
	(1,2,3,10)	.022	.036	.040
	(1,3,5,7)	.033	.047	.053
	(2.67,4,4,5.33)	.033	.045	.059
	(4,4,4,4)	.043	.043	.046
	(5.33,4,4,2.67)	.074	.050	.051
	(7,5,3,1)	.134	.059	.050
	(10,3,2,1)	.173	.075	.063
	(13,1,1,1)	.236	.076	.055

TABLE 5. Estimated Empirical Power of the Tests, Nominal Size = .05, Sample Size Condition (8,8,8,8), Mean Structure ($\mu_3 < \mu_1 = \mu_2 < \mu_4$).

Noncentrality Parameter	Variance Conditions (7,5,3,1) ANOVA F	Box	Welch	(1,3,5,7) ANOVA F	Box	Welch
0.0	.075	.038	.050	.069	.025	.031
0.6	.138	.088	.213	.131	.113	.088
1.0	.313	.225	.531	.306	.250	.200
1.3	.538	.438	.775	.488	.425	.325
1.6	.756	.656	.931	.681	.581	.469
2.0	.950	.831	.999	.856	.788	.700

TABLE 6. Estimated Empirical Power of the Tests, Nominal Size = .05, Sample Size condition (8,8,8,8), Mean Structure ($\mu_1 < \mu_2 < \mu_3 < \mu_4$).

Noncentrality Parameter	Variance Conditions (1,1,1,13) or (13,1,1,1) ANOVA F	Box	Welch	(4,4,4,4) ANOVA F	Box	Welch
0.0	.119	.088	.063	.088	.031	.050
0.6	.075	.100	.200	.144	.113	.125
1.0	.319	.200	.500	.344	.300	.300
1.3	.463	.300	.738	.513	.481	.481
1.6	.631	.425	.894	.688	.663	.644
2.0	.825	.625	.999	.900	.875	.850

TABLE 7. Variance Conditions of the Study.

Coefficient of Variation of the Variances	Variance Condition	Bias (b) Values Sample Size Conditions (6,8,9,9)	(5,7,10,14)
1.299	(1,1,1,13)	0.875	0.586
0.884	(1,2,3,10)	0.884	0.663
0.559	(1,2,3,7)	0.894	0.754
0.235	(2.67,4,4,5.33)	0.956	0.889
0.0	(4,4,4,4)	1.0	1.0
0.235	(5.33,4,4,2.67)	1.048	1.134
0.559	(7,5,3,1)	1.127	1.397
0.884	(10,3,2,1,)	1.231	1.568
1.299	(13,1,1,1)	1.352	1.778

TABLE 8. Estimated Empirical Power of the Tests, Nominal Size = .05, Sample Size Condition (5,7,10,14), Mean Sturcture ($\mu_3 < \mu_1 = \mu_2 < \mu_4$).

Noncentrality Parameter	Variance Condition (1,3,5,7) ANOVA F	Box	Welch
0.0	.019	.038	.062
0.6	.100	.163	.125
1.0	.319	.400	.300
1.3	.525	.619	.463
1.6	.725	.800	.625
2.0	.913	.931	.813

TABLE 9. Estimated Empirical Power of the Tests, Nominal Size = .05 Sample Size Condition (6,8,9,9), Mean Structure ($\mu_1 < \mu_3 = \mu_4 < \mu_2$).

Noncentrality Parameter	Variance Conditions (7,5,3,1) Box	Welch	(13,1,1,1) Box	Welch
0.0	.031	.075	.075	.056
0.6	.150	.113	.125	.225
1.0	.225	.200	.181	.419
1.3	.375	.306	.275	.638
1.6	.500	.425	.350	.825
2.0	.700	.600	.538	.981

J. B. Dijkstra and P. S. P. J. Werter (1981)

Purpose and Method

Dijkstra and Werter compared the size and power of three tests, the second order test of James, the Welch test and the test of Brown-Forsythe. Although James' second order method was ruled out as an alternative to the ANOVA F statistic in the Alternative Methodologies section, its values of size and power were listed to demonstrate its similarities to Welch's test. The size study included three groups, four groups and six groups. The sample sizes ranged from four to twenty and the standard deviations ranged from one to three. The power study included two sets of four groups, one equal the other unequal. Three different alternatives were used with the same standard deviations as the size study. For each set of criterion 10,000 replications were simulated.

Size

The range of the size of all three test statistics for the various combinations was from 3.5% to 7.5%, so all three have excellent control of the size as demonstrated in TABLE 10. The second order test of James was the test statistic that remained closest to the nominal value in nearly all cases whereas Welch's test and the test of Brown-Forsythe behaved very similarly but both were slightly higher than the nominal value.

Power

Uniformly none of the tests are more powerful than the other two. The test of Welch and the second order James' test are almost identical differing only in the third decimal place. As was found

earlier by Brown and Forsythe, when an extreme mean was paired with the largest variance the Brown-Forsythe test had superior power by as much as 24%. Conversely, when the extreme mean was paired with the smallest variance Welch's test and the second order test of James had superior power by as much as 32%. When the alternative included two extreme means (5,0,0,0.5) or (0.5,0,0,5) paired with the extreme variances (1,2,2,3) the test with superior power was dictated by whether the largest extreme mean coincided with the largest variance or not. Thus for this alternative the behavior was exactly the same as above when there was only one extreme mean, but with equal sample sizes the difference in power was around 21% and with unequal sample sizes the difference was only as high as 6%. Results of the power study are displayed in TABLE 11.

Comments

The sample sizes chosen for the size study were adequate but a few more variance combinations would have been helpful for both the size study and the power study. A variance combination with just one extreme variance (i.e. (1,1,1,3)) would have added to the utility of this investigation. Some cases of the noncentrality parameter were chosen poorly (i.e. (5,0,0,0.5)). It would have been better to choose an alternative like (1,0,0,0.5) or (0.5,0,0,0.5). The detection of a difference in the means is more difficult when the means are closely grouped. Thus if a test has good power when there is only a small difference in the means then it would follow logically that the power would be even higher when the means differ by a large amount. For the above alternative the power of all three tests is so high that it is

difficult to assess which test is superior. However, the other alternatives do show more clearly the power behavior of the tests. It might have been helpful to include the size and power of the ANOVA F along with the other tests as well. Because the behavior of the Welch test and the second order test of James are nearly identical the simpler test by Welch would be the test of choice because of the complexity in calculating James' second order test. The findings of this paper are consistent with the findings of the previous papers. New findings include the similarity of Welch's test and James' test and the power behavior under the (5,0,0,0.5) and (0.5,0,0,5) alternatives.

TABLE 10. Empirical Type I Error Probabilities, Nominal size = .05.

Sample Size Condition	Standard Deviation Condition	Brown-Forsythe	James	Welch
(4,4,4)	(1,1,1)	.037	.044	.041
	(1,2,3)	.048	.053	.049
(4,6,8)	(1,1,1)	.044	.050	.049
	(1,2,3)	.050	.043	.042
(4,4,4,4)	(1,1,1,1)	.040	.052	.050
	(1,2,2,3)	.044	.057	.056
(4,6,8,10)	(1,1,1,1)	.044	.049	.052
	(1,2,2,3)	.052	.044	.045
	(3,2,2,1)	.059	.054	.059
(10,10,10,10)	(1,1,1,1)	.048	.050	.051
	(1,2,2,3)	.056	.050	.051
(10,14,16,20)	(1,1,1,1)	.046	.046	.046
	(1,1.5,2,3)	.061	.050	.050
	(3,2,1.5,1)	.062	.046	.048
(4,4,4,4,4,4)	(1,1,1,1,1,1)	.035	.053	.062
	(1,1,2,2,3,3)	.046	.064	.075
(4,6,8,10,12,14)	(1,1,1,1,1,1)	.043	.053	.062
	(1,1,2,2,3,3)	.065	.051	.056
	(3,3,2,2,1,1)	.057	.058	.069
(10,10,10,10,10,10)	(1,1,1,1,1,1)	.048	.050	.052
	(1,1,2,2,3,3)	.065	.051	.053
(10,10,15,15,20,20)	(1,1,2,2,3,3)	.068	.051	.053
	(3,3,2,2,1,1)	.063	.053	.056

TABLE 11. Empirical Power of the Tests, Nominal Size = .05.

Sample Size Condition	Standard Deviation Condition	Mean Structure	Brown-Forsythe	James	Welch
(4,4,4,4)	(1,1,1,1)	(0,0,0,0)	.035	.053	.052
		(5,0,0,0.5)	1.000	.999	.999
		(3,0,0,0)	.951	.875	.874
(4,4,4,4)	(1,2,2,3)	(0,0,0,0)	.046	.057	.057
		(3,0,0,0)	.312	.616	.616
		(0,0,0,3)	.306	.220	.220
		(5,0,0,0.5)	.751	.974	.972
		(0.5,0,0,5)	.660	.449	.447
(4,6,8,10)	(1,1,1,1)	(0,0,0,0)	.047	.052	.055
		(3,0,0,0)	.986	.935	.941
		(0,0,0,3)	1.000	1.000	1.000
(4,6,8,10)	(1,2,2,3)	(0,0,0,0)	.057	.050	.051
		(3,0,0,0)	.556	.873	.876
		(0,0,0,3)	.746	.521	.524
		(5,0,0,0.5)	.983	.999	.999
		(0.5,0,0,5)	.999	.923	.926

R. R. Wilcox, V. L. Charlin and K. L. Thompson (1986)

Purpose and Method

The authors of this paper compare the performance, in terms of size and power, of the ANOVA F statistic, Welch's test and the test of Brown-Forsythe. The size study included four groups and six groups with sample sizes ranging from four to fifty. The standard deviations ranged from one to four and standard deviation conditions included one extreme standard deviation, equally spaced standard deviations, as well as other combinations. The power study included the same sample size and standard deviation combinations with one alternative hypothesis. For each set of criterion 10,000 replications were simulated.

Size

With as many as fifty observations per group and equal sample sizes in four groups the ANOVA F can be very unsatisfactory when there is one extreme variance. Comparing the equal sample size cases (11,11,11,11), (21,21,21,21) and (50,50,50,50) with one extreme standard deviation (1,1,1,4) it becomes apparent that the robustness of the ANOVA F improves very slowly as the sample sizes increase, and it is not obvious when, if ever, the sample sizes would be large enough to indicate that the ANOVA F would be acceptable in terms of its size. As TABLE 12 shows the empirical probability of a Type I error starts at 11% and only reduces to 9% when the groups have fifty observations. For equal sample sizes Welch's test performs better than the test of Brown-Forsythe in the sense that the maximum

empirical Type I error probability for the Brown-Forsythe test was 8.4% while the maximum value for Welch's test was only 6.0%.

However, when the four groups had unequal sample sizes the choice between the Brown-Forsythe test and Welch's test is not as clear-cut. There are instances where one will have a slight advantage over the other but both tests have Type I error rates between 4.4% and 8.6%. When the larger standard deviations were paired with the larger sample sizes Welch's test remained closer to the nominal value. Conversely, when the larger standard deviations were paired with the smaller sample sizes the Brown-Forsythe test remained closer to the nominal value. The Type I error rate for the ANOVA F ranged from 2.7% to 27.9%, being very conservative when the larger standard deviations coincided with the larger sample sizes and quite liberal when larger standard deviations coincided with the smaller sample sizes.

With six groups of equal sample size Welch's test had a slight edge over the Brown-Forsythe. When the group sizes were unequal the behavior of all three tests were very similar to the four group situations. Overall, the edge would be given to the Welch test because of its smaller maximum empirical rate of 8.6% as compared to the maximum empirical rate of 10% for the Brown-Forsythe test.

Power

Even when the variances are equal there is only a slight loss of power, around 1-2%, when using Welch's test or the Brown-Forsythe test as compared to using the ANOVA F. As several authors have noted previously, when extreme means are paired with small standard deviations Welch's test has superior power by as much as 69%, as shown

in TABLE 13. When extreme means are paired with larger standard deviations the Brown-Forsythe test has superior power but only by as much as 26%. In the cases where the Welch test does not have superior power the ANOVA F has about 20-25% higher power than Welch's test but for these exact cases the ANOVA F has particularly poor control over the size so that the increase in power may be due to the inflated Type I error rate.

Comments

The combination of sample sizes and standard deviations were adequate to show the size behavior of the various tests. The findings of this paper are consistent with those previously reviewed. This paper uncovered a condition when the performance of the Welch test is not good. This occurs when there is one or more extreme standard deviations and they are paired with the smaller sample sizes the empirical Type I error rate of Welch's test reaches its maximum value, around 8%. Other tests perform much worse than Welch's test and don't reveal any consistent pattern. Although the combination of sample sizes and standard deviations were adequate for this investigation the power performances were limited by the use of only one alternative hypothesis. Other alternatives would have been helpful in this study.

TABLE 12. Empirical Type I Error Probabilities, Nominal size = .05.

Sample Size Condition	Standard Deviation Condition	ANOVA F	Brown-Forsythe	Welch
(11,11,11,11)	(1,1,1,1)	.048	.046	.055
	(1,2,3,4)	.068	.061	.060
	(4,1,1,1)	.109	.084	.055
(21,21,21,21)	(1,1,1,1)	.051	.050	.056
	(1,2,3,4)	.069	.065	.056
	(4,1,1,1)	.097	.084	.055
(50,50,50,50)	(4,1,1,1)	.088	.084	.044
(4,8,10,12)	(1,1,1,1)	.051	.048	.072
	(4,3,2,1)	.173	.065	.086
	(1,1,1,4)	.041	.075	.069
	(4,1,1,1)	.279	.081	.082
(6,10,16,20)	(1,1,1,1)	.053	.069	.065
	(4,3,2,1)	.194	.059	.070
	(1,1,1,4)	.027	.077	.062
	(4,1,1,1)	.275	.072	.068
(15,15,15,15,15,15)	(1,1,1,1,1,1)	.049	.048	.062
	(1,1,1,4,4,4)	.080	.071	.064
	(1,1,1,1,1,4)	.119	.095	.064
(6,10,15,18,21,25)	(1,1,1,1,1,1)	.047	.047	.075
	(1,1,1,4,4,4)	.029	.080	.068
	(4,4,4,1,1,1)	.234	.069	.080
	(1,1,1,1,1,4)	.041	.100	.073
	(4,1,1,1,1,1)	.309	.091	.078
	(4,3,3,1,1,4)	.091	.062	.080

TABLE 13. Empirical Power of the Tests, Nominal size = .05, Mean Structure for All Cases (1.2,0,0,0) or (1.2,0,0,0,0,0).

Sample Size Condition	Standard Deviation Condition	ANOVA F	Brown-Forsythe	Welch
(11,11,11,11)	(1,1,1,1)	.794	.789	.773
	(4,1,1,1)	.244	.206	.118
(21,21,21,21)	(1,1,1,1)	.983	.983	.981
	(4,1,1,1)	.372	.348	.180
(50,50,50,50)	(4,1,1,1)	.604	.593	.334
(4,8,10,12)	(1,1,1,1)	.396	.366	.412
	(4,3,2,1)	.232	.094	.109
	(1,1,1,4)	.060	.112	.392
	(4,1,1,1)	.359	.114	.106
(6,10,16,20)	(1,1,1,1)	.592	.564	.570
	(4,3,2,1)	.282	.111	.107
	(1,1,1,4)	.050	.144	.545
	(4,1,1,1)	.381	.132	.108
(15,15,15,15,15,15)	(1,1,1,1,1,1)	.908	.907	.898
	(1,1,1,4,4,4)	.152	.137	.825
	(1,1,1,1,1,4)	.334	.275	.883
(6,10,15,18,21,25)	(1,1,1,1,1,1)	.546	.525	.552
	(1,1,1,4,4,4)	.038	.104	.505
	(4,4,4,1,1,1)	.307	.107	.113
	(1,1,1,1,1,4)	.070	.168	.546
	(4,1,1,1,1,1)	.422	.154	.113
	(4,3,3,1,1,4)	.154	.108	.113

A. J. Tomarken and R. C. Serlin (1986)

Purpose and Method

Tomerken and Serlin study the size and power of the ANOVA F, Welch, Brown-Forsythe, Kruskal-Wallis and inverse normal scores tests. Only the first three tests will be discussed in this investigation. The size and power studies included three groups and four groups with sample sizes ranging from six to thirty. For the power study four alternatives were investigated with four different sample size and variance combinations:equal variances, equal pairing (equal sample sizes and increasing variances), direct pairing (increasing sample sizes and increasing variances), and inverse pairing (increasing sample sizes and decreasing variances). For each set of criterion 1,000 replications were simulated. Tables of the results from this investigation were averaged across cases that were similar in design. The design structure is shown in TABLE 14, only cases with the same letter were averaged together.

Size

With three groups of equal sample sizes, when the variances were unequal, in three of the four cases the empirical rejection rates of the ANOVA F statistic exceeded the robustness criterion of 7% adopted by the authors. With four groups of equal sample sizes and unequal variances the ANOVA F performed more acceptably, only exceeding the robustness criterion once in the four cases, but the average was higher than that of Welch's test or Brown-Forsythe's test. Both Welch's test and Brown-Forsyth's test performed adequately in the

equal sample size cases with Welch's test remaining slightly closer to the nominal value.

In the unequal sample size cases the ANOVA F showed marked deviations, being too conservative when larger sample sizes were paired with larger variances and too liberal when larger sample sizes were paired with smaller variances. Results are shown in TABLE 15. The Welch test and the Brown-Forsythe test remained within the authors' robustness limits in both the three and four group cases. Once again the Welch test remained slightly closer to the nominal level than the Brown-Forsythe test. As noted in the previous paper, when larger sample sizes were paired with smaller variances Welch's test had a slightly poorer performance. Overall, Welch's test showed the best control of the Type I error rate.

Power

Results summarizing the behavior of the various tests when the variances are equal are shown in TABLE 16. For all cases mean structures were specified to an estimated ANOVA power of 0.70 and nominal level 0.05. Additional mean structures were specified with estimated ANOVA power of 0.85 and 0.55, these cases are denoted by * and ** respectively. As expected the ANOVA F had the highest power but only by about 1-5%. Brown-Forsythe's test had slightly higher power than Welch's test. It may be suprising that only minimal losses in power are incurred when the ANOVA F alternatives are used in the equal variance cases.

The results from the three sample size and variance combinations, equal pairing, direct pairing and inverse pairing are shown in TABLE 17. Although for the equal sample size cases the ANOVA F was the most powerful when the extreme mean was paired with the largest variance, the Type I error assessments showed that it was frequently too liberal under these conditions, particularly when there were three groups. A striking consistency existed across all the cases studied. In each of the three sample size and variance combinations, the Welch test proved to be the most powerful procedure when means were equally spaced apart, when extreme means were paired with the smallest variances and when two identical means were situated midway between two extreme means. Although the Welch test was consistently superior for these mean patterns, its relative advantage varied somewhat across conditions. Its rejection rates ranged from 5% to 15% higher than the Brown-Forsythe test when the means were equally spaced apart. When extreme means were paired with the smallest variances Welch's superiority ranged from 5% to 35% and when two identical means were situated between two extreme means the superiority ranged from 9% to 21%.

Although the Welch test was unequivocally the test of choice for three of the four mean structures, the Brown-Forsythe test was optimal, though less clearly so, for the mean structure where an extreme mean was paired with the largest variance. The superiority of Brown-Forsythe's test ranged from 8% to 18%. Even though the ANOVA F had the highest power its severe lack of control of the size renders it an unreasonable test.

Comments

This study included sufficient combinations of sample sizes and variances to see the size and power behavior of the tests. The only variance condition that was omitted that may have been helpful would have been the situation where just one extreme variance existed. Although the figures in the tables accurately described the behavior of the various tests, the way they were tabulated was confusing. Even though there would have been four times as many tables if listed separately the authors could have chosen the ones that were outstanding or demonstrated a point that was being made. The findings in this paper were consistent with previous papers. New findings included the demonstration that the same behavior of the tests found earlier could be expected with medium and large sample sizes and confirmation that the ANOVA F fails to adequately control the Type I error rate even when the sample sizes large.

TABLE 14. Design of the Monte Carlo Investigation.

Sample Size Condition	Variance Conditons (6,6,6)	(12,4,1)	(6,2,1)	(1,4,12)	(1,2,6)
(20,20,20)	A	B	B		
(12,12,12)	A	B	B		
(30,20,10)	A	C	C	D	D
(18,12,6)	A	C	C	D	D

	Variance Conditions (6,6,6,6)	(12,6,4,1)	(6,3,2,1)	(1,4,6,12)	(1,2,3,6)
(20,20,20,20)	A	B	B		
(12,12,12,12)	A	B	B		
(30,24,16,10)	A	C	C	D	D
(18,14,10,6)	A	C	C	D	D

TABLE 15. Empirical Type I Error Probabilities, Nominal Size = .05, Probabilities Given are Averaged Across Cases With Same Letter.

Sample Size Condition	Variance Condition	ANOVA F	Welch	Brown-Forsythe
(All 4 cases)	$(\sigma_1^2=\sigma_2^2=\sigma_3^2)$	.053(0)	.050(0)	.051(0)
$(n_1=n_2=n_3)$	$(\sigma_1^2>\sigma_2^2>\sigma_3^2)$	.069(3)	.048(0)	.062(0)
$(n_1>n_2>n_3)$	$(\sigma_1^2>\sigma_2^2>\sigma_3^2)$	.022(0)	.049(0)	.061(0)
$(n_1>n_2>n_3)$	$(\sigma_1^2<\sigma_2^2<\sigma_3^2)$	.167(4)	.057(0)	.064(1)
(All 4 cases)	$(\sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_4^2)$	.053(0)	.055(0)	.052(0)
$(n_1=n_2=n_3=n_4)$	$(\sigma_1^2>\sigma_2^2>\sigma_3^2>\sigma_4^2)$	.064(1)	.048(0)	.059(0)
$(n_1>n_2>n_3>n_4)$	$(\sigma_1^2>\sigma_2^2>\sigma_3^2>\sigma_4^2)$	.025(0)	.050(0)	.059(0)
$(n_1>n_2>n_3>n_4)$	$(\sigma_1^2<\sigma_2^2<\sigma_3^2<\sigma_4^2)$	.144(4)	.056(0)	.059(0)

NOTE: The number in parentheses is the number of times out of the four cases that the empirical rejection rate was greater than .07.

TABLE 16. Empirical Power of the Tests, Nominal Level = .05, Equal Variances, Probabilities Given are Averaged Across Cases With Same Letter.

Sample Size Condition	Mean Structure	Number Of Cases	ANOVA F	Welch	Brown-Forsythe
(All 4 cases)	$(\mu_1>\mu_2>\mu_3)$	4	.697	.665	.684
$(n_1=n_2=n_3)$	$(\mu_1>\mu_2=\mu_3)$	2	.712	.693	.709
$(n_1>n_2>n_3)$	$(\mu_1>\mu_2=\mu_3)$	2	.700	.662	.674
$(n_1>n_2>n_3)$	$(\mu_1=\mu_2>\mu_3)$	2	.683	.639	.664
(All 4 cases)	$(\mu_1>\mu_2>\mu_3>\mu_4)$	4	.698	.664	.684
$(n_1=n_2=n_3=n_4)$	$(\mu_1>\mu_2=\mu_3=\mu_4)$	2	.704	.672	.702
$(n_1>n_2>n_3>n_4)$	$(\mu_1>\mu_2=\mu_3=\mu_4)$	2	.682	.628	.666
$(n_1>n_2>n_3>n_4)$	$(\mu_1=\mu_2=\mu_3>\mu_4)$	2	.684	.637	.667
(All 4 cases)	$(\mu_1>\mu_2=\mu_3>\mu_4)$	4	.699	.660	.685

TABLE 17. Empirical Power of the Tests, Nominal Level = .05, For All Cases Mean Structures were Specified to an Estimated ANOVA Power of .700.

Sample Size and Variance Condition	Mean Structure	Number Of Cases	ANOVA F	Welch	Brown-Forsythe
Equal Pairing	$\mu_1>\mu_2>\mu_3$	4	.683	.765	.659
$(n_1=n_2=n_3)$	$\mu_1>\mu_2=\mu_3$	4	.646	.493	.628
$(\sigma_1^2>\sigma_2^2>\sigma_3^2)$	$(\mu_1>\mu_2=\mu_3)^*$	(2)	(.803)	(.645)	(.784)
	$\mu_1=\mu_2>\mu_3$	4	.766	.938	.743
	$(\mu_1=\mu_2>\mu_3)^{**}$	(1)	(.554)	(.864)	(.529)
Direct Pairing	$\mu_1>\mu_2>\mu_3$	4	.665	.909	.855
$(n_1>n_2>n_3)$	$\mu_1>\mu_2=\mu_3$	4	.634	.680	.804
$(\sigma_1^2>\sigma_2^2>\sigma_3^2)$	$\mu_1=\mu_2>\mu_3$	4	.760	.992	.940
	$(\mu_1=\mu_2>\mu_3)^{**}$	(3)	(.547)	(.945)	(.803)
Inverse Pairing	$\mu_1>\mu_2>\mu_3$	4	.691	.538	.387
$(n_1>n_2>n_3)$	$(\mu_1>\mu_2>\mu_3)^*$	(2)	(.840)	(.701)	(.514)
$(\sigma_1^2<\sigma_2^2<\sigma_3^2)$	$\mu_1>\mu_2=\mu_3$	4	.767	.722	.421
	$\mu_1=\mu_2>\mu_3$	4	.646	.298	.384
	$(\mu_1=\mu_2>\mu_3)^*$	(4)	(.756)	(.404)	(.514)

* Additional power assessment with estimated ANOVA power of .850.

** Additional power assessment with estimated ANOVA power of .550.

TABLE 17 cont.

Sample Size and Variance Condition	Mean Structure	Number Of Cases	ANOVA F	Welch	Brown-Forsythe
Equal Pairing	$\mu_1>\mu_2>\mu_3>\mu_4$	4	.681	.779	.660
$(n_1=n_2=n_3=n_4)$	$\mu_1>\mu_2=\mu_3=\mu_4$	4	.634	.444	.622
$(\sigma_1^2>\sigma_2^2>\sigma_3^2>\sigma_4^2)$	$(\mu_1>\mu_2=\mu_3=\mu_4)^*$	(4)	(.756)	(.570)	(.730)
	$\mu_1=\mu_2=\mu_3>\mu_4$	4	.770	.961	.751
	$(\mu_1>\mu_2=\mu_3>\mu_4)^{**}$	(2)	(.584)	(.924)	(.564)
	$\mu_1>\mu_2=\mu_3>\mu_4$	4	.688	.816	.667
Direct Pairing	$\mu_1>\mu_2>\mu_3>\mu_4$	4	.665	.900	.817
$(n_1>n_2>n_3>n_4)$	$\mu_1>\mu_2=\mu_3=\mu_4$	4	.613	.567	.756
$(\sigma_1^2>\sigma_2^2>\sigma_3^2>\sigma_4^2)$	$\mu_1=\mu_2=\mu_3>\mu_4$	4	.770	.994	.925
	$(\mu_1=\mu_2=\mu_3>\mu_4)^{**}$	(4)	(.550)	(.964)	(.777)
	$\mu_1>\mu_2=\mu_3>\mu_4$	4	.637	.886	.799
Inverse Pairing	$\mu_1>\mu_2>\mu_3>\mu_4$	4	.712	.615	.467
$(n_1>n_2>n_3>n_4)$	$\mu_1>\mu_2=\mu_3=\mu_4$	4	.792	.848	.497
$(\sigma_1^2<\sigma_2^2<\sigma_3^2<\sigma_4^2)$	$\mu_1=\mu_2=\mu_3>\mu_4$	4	.645	.301	.430
	$(\mu_1=\mu_2=\mu_3>\mu_4)^*$	(4)	(.754)	(.402)	(.558)
	$\mu_1>\mu_2=\mu_3>\mu_4$	4	.728	.678	.471

* Additional power assessment with estimated ANOVA power of .850.

** Additional Power assessment with estimated ANOVA power of .550.

CONCLUSIONS AND RECOMMENDATIONS

An unexpected finding was the lack of robustness of the ANOVA F when the variances were unequal. Even when the sample sizes were equal the empirical Type I error probabilities could be twice as high as desired, particularly when there was only one extreme variance. As seems to be well known, when the sample sizes were unequal the ANOVA F showed marked deviations from the nominal level, being too conservative when larger variances were paired with larger sample sizes and too liberal when larger variances were paired with smaller sample sizes. All other tests did remarkably well in terms of controlling the Type I error rate. The test that had the best control of the Type I error rate was Welch's test. Even at its worst the empirical rate rarely reached as high as 8%. This occurred when there was one extreme variance and when the sample sizes and variances were inversely paired.

Although only one paper considered the test of Box, the paper did raise some doubts about the utility of the test. It was not the test of choice even when the sample sizes were equal and rarely had superior power over Welch's test.

The first order test of James was similar to Welch's test in terms of power but did not control the Type I error rate as well. James' second order test was equivalent to Welch's test in terms of size and power but, because of its complexity in calculation it was not considered better than Welch's test.

This leaves the choice of an alternative to the ANOVA F between the Brown-Forsythe test and the Welch test. Another surprising result was that there was only about a 1% to 5% loss in power as compared to the ANOVA F when using these two tests when variances were equal. Unequivocally it cannot be said that one test is consistently superior to the other. The Brown-Forsythe test has superior power only when extreme means coincide with the largest variances. In most other cases Welch's test has superior power. For example, when extreme means are paired with the smallest variances or when means are equally spaced apart Welch's test has superior power. When two identical means are situated midway between two extreme means Welch's test once again has superior power. Thus for an experimenter wanting to perform an omnibus test, Welch's test should be used. There is little doubt that in some situations Welch's test will not be the most powerful but on these occasions the loss in power would only be 1% to 25%.

Another approach for selecting an alternative procedure would be to carefully investigate the experimental data and choose the test that would provide the most power, either Welch's test or Brown-Forsythe's test. The researcher could also perform both tests and go with the alternative hypothesis if either test rejects the null hypothesis.

Further research possibilities might include the investigation of the size and power of the alternative tests when the variances are unequal as well as the observations lacking normality. Another extension of this report could be the investigation of higher order treatment structures when the variances are unequal.

Typical experimental situations where Welch's test would be particularly beneficial could be experiments involving preservatives in meat products or experiments involving yields of grains. In preservative experiments a situation commonly referred to as masking occurs. Masking is where a difference between two means may not be detected because of a large variance in a third mean. This frequently occurs because the controls used in the experiment typically have large variances. An example of masking might be a situation with a sample size condition of (6,10,16,20), a standard deviation condition of (1,1,1,4) and a mean structure of (1.2,0,0,0). In this particular situation Welch' test is 40% (54-14) more powerful than Brown-Forsythe's test. With a sample size condition of (4,8,10,12) and the same standard deviation condition and mean structure as above a 28% (39-11) gain in power is acheived by using Welch's test. Even when there are six groups, see TABLE 13, the superiority in power of Welch's test can be as high as 40%.

For the grain yield experiment a common situation occurs where low yields have small variances and high yields have large variances. This is a direct pairing situation. An example of this direct pairing might be the following: sample size condition $(n_1 > n_2 > n_3 > n_4)$, variance condition $(\sigma_1^2 > \sigma_2^2 > \sigma_3^2 > \sigma_4^2)$, and mean structure $(\mu_1 > \mu_2 > \mu_3 > \mu_4)$. For this situation the Welch test had a power of 90% and the Brown-Forsythe test had a power of 82%. Other examples of this direct pairing are shown in TABLE 17.

These are just a few of the typical situations where an alternative to the ANOVA F such as the Welch test could be very beneficial to researchers. In general, whether the sample sizes are equal or not, if the variances are unequal the Welch test should be used because of its excellent control of the Type I error rate and usually superior power.

LITERATURE CITED

Bishop, T. A., and Dudewicz, E. J. (1978). "Exact analysis of variance with unequal variances". Technometrics 20:419-430.

Box, G. E. P. (1954). "Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification". Annals of Mathematical Statistics 25:290-302.

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

Dijkstra, J. B., and Werter, P. S. P. J. (1981). "Testing the equality of several means when the population variances are equal". Communications in Statistics. B. Simulation and Computation 10:557-569.

James, G. S. (1951). "The comparison of several groups of observations when the ratios of the population variances are unknown". Biometrika 38:324-329.

Kohr, R. L., and Games, P. A. (1974). "Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances". Journal of Experimental Education 43:61-69.

Levy, K. J. (1978a). "Some empirical power results associated with Welch's robust analysis of variance technique". Journal of Statistical Computing and Simulation 8:43-48.

Milliken, G. A., and Johnson, D. E. (1984). Analysis of Messy Data. New York: Van Nostrand Reinhold Company.

Satterwaite, F. E. (1941). "Synthesis of variance". Psychometrika 6:309-316.

Scheffe', H. (1944). "A note on the Behrens-Fisher problem". Annals of Mathematical Statistics 15:430-432.

Stein, C. (1945). "A two-sample test for a linear hypothesis whose power is independent of the variance". Annals of Mathematical Statistics 16:243-258.

Tomarken, A. J., and Serlin, R. C. (1986). "Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures'. Quantitative Methods in Psychology 99:90-99.

Wang, Y. Y. (1971). "Probabilities of the Type I errors of the Welch tests for the Behrens-Fisher problem". Journal of the American Statistical Association 66:605-608.

Welch, B. L. (1936). "Specification of rules for rejecting too variable a product, with particular reference to an electric lamp problem". Journal of the Royal Statistical Society, Supp. 3 26-48.

Welch, B. L. (1951). "On the comparison of several mean values: An alternative approach". Biometrika 38:330-336.

Wilcox, R. R. (1987). "A heteroscedastic ANOVA procedure with specified power". Journal of Educational Statistics 12:271-281.

Wilcox, R. R., Charlin, V. L., and Thompson, K. L. (1986). "New monte carlo results on the robustness of the ANOVA F, W and F^* statistics". Communications in Statistics, B, Simulation and Computation 15:933-943.

AN INVESTICATION OF THE POWER OF VARIOUS ALTERNATIVES TO THE ANOVA F STATISTIC WHEN POPULATION VARIANCES ARE UNEQUAL

by

Mark Allan Sorell

B. S., Kansas State University, 1986

AN ABSTRACT OF A MASTER'S REPORT

Submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1988

ABSTRACT

Alternative procedures for testing the equality of population means when the assumption of equal variances is violated are discussed. A numerical example is illustrated for each alternative procedure. Five papers which compare the performance of these alternatives are reviewed. The tests are appraised in terms of significance level and power. Whereas other procedures had limitations in several contexts, the findings of this report indicate the test by Welch is the test of choice because of its excellent control of the Type I error rate and usually superior power.