The University of Manitoba

The Graduate School

Department of Psychology

Procedures for the Sequential Testing of Variance Homogeneity and Equality of Means in the One-way Analysis of Variance Fixed Effects Model

A Thesis in

Quantitative Psychology

by

Jennifer J. Clinch

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

PROCEDURES FOR THE SEQUENTIAL TESTING OF VARIANCE HOMOGENEITY AND EQUALITY OF MEANS IN THE ONE-WAY ANALYSIS OF VARIANCE FIXED EFFECTS MODEL

BY

JENNIFER J. CLINCH

A dissertation submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF ARTS

Acknowledgments

I would like to thank Dr. Harvey Keselman, Dr. Larry Breen, Dr. Brian Macpherson and Dr. Roy Gabriel for their participation in the preparation of this thesis.

My thanks are especially due to Harvey Keselman for his help and patience as my advisor.

Dr. Brian Macpherson found the time for several helpful discussions on transformations and on the Box-Andersen variance test, which was much appreciated.

Finally, I would like to express my gratitude to Cynthia Jordan for her support and encouragement during the writing of the results section.

Table of Contents

Table of Contents (Cont'd)

List of Tables

List of Tables (Cont'd)

List of Figures

List of Figures (Cont'd)

Abstract

F-statistics calculated from an Analysis of Variance (ANOVA) are biased when the treatment populations sampled have unequal variances, and especially so when the samples are unequal in size (Glass, Peckham, and Sanders, 1972). Several solutions to this problem have been developed which use the standard F tables but adjust the calculated statistics and their degrees of freedom by factors related to the heterogeneous variances (e.g., Welch, 1951). While these alternative procedures control the probability of Type I error when the ANOVA assumption of homogeneity of variance is violated, they may have less power than the ANOVA when the assumptions are met (Kohr and Games, 1974). Prior testing of the validity of the variance homogeneity assumption would allow a choice of tests and hopefully optimize control of Type I error and power, but this procedure has not been popular because traditional variance tests are highly sensitive to non-normality, while the ANOVA F-test of mean differences is not (Box, 1953). Using more recently developed robust variance tests (e.g., Brown and Forsythe, 1974b), the present research re-examined the question of allowing the outcome of a test of variance homogeneity to dictate the choice of a test on means.

Monte Carlo methods were used to simulate a one-way fixed effects ANOVA design having four treatment groups. The performance of individual means tests and several sequential procedures (in which a variance test chose between the ANOVA and one of two alternate procedures) was evaluated under a variety of conditions representing all possible combinations

of: a) four degrees of variance heterogeneity plus equal variances, b) three degrees of group size inequality plus equal sizes, c) positive and negative pairing of variances and group sizes, d) two population shapes (normal and chi-square), e) two levels of overall sample size, f) two patterns of group mean differences plus equal means, and g) positive and negative pairing of group means and variances. The means tests evaluated were a) the ANOVA F-test, b) Welch's (1951) procedure, and c) Brown and Forsythe's (1974a) F* test, while the variance tests used to choose between the ANOVA F-test and one of the other two tests were a) Bartlett's (1937) test, b) Box and Andersen's (1955) test, c) Brown and Forsythe's (1974b) test on absolute deviations from the median, d) Miller's (1968) jacknife test, and e) the Box-Scheffé test (Box, 1953; Scheffé, 1959). Data transformations were also performed in order to assess this procedure as a method of removing variance heterogeneity; subsequent F-tests on transformed data were compared with the other means tests.

In accordance with previous studies (e.g., Kohr and Games, 1974) the results demonstrated the superior robustness and power of the Welch test when the population sampled was normal, however in the chi-square population it failed to control Type I error rates. Brown and Forsythe's F* test was the preferred test in this latter situation and was therefore recommended overall since its Type I error rates were still acceptable even on those occasions when the Welch test was more robust. Although the power of the F* test was invariably less than that of the Welch test, it usually was close to a priori power calculated for the ANOVA F-test.

When variance tests were used to choose between tests of mean

equality, control of Type I error rates was never substantially better and usually worse than when uniformly adopting the alternate test to the ANOVA F-test: this was especially true for the situations where the ANOVA F-test was positively biased. On those few occasions when Type I error control was better for the sequential procedures, power was considerably less than when uniformly adopting the alternate test: the small gain in improved Type I error rates was not considered worth the cost in power. Thus Brown and Forsythe's F* test was considered better overall than all single or sequential testing procedures.

Finally, F-tests on transformed data did not perform as well as those on the original untransformed data.

Introduction

Scheffé (1970, p.1501) states "The most commonly occurring problem in applied statistics is, in my opinion, the comparison of the means of two populations." Usually this comparison is made using Students t-test. Probably the second most frequently occurring problem is the comparison of the means of more than two populations, for which the usual method is the analysis of variance (ANOVA) F-test. Both the t- and F-test have been derived on the basis of certain simplifying assumptions, which should be met before these tests are used. These assumptions are: (a) Observations are sampled from normally distributed populations. (b) Each population sampled has the same variance, i.e., variances are homogeneous. (c) Errors associated with any pair of observations are independent. Failure to meet any of these assumptions may result in loss of accuracy of the statistical test. When inferences are made, based on the results of statistical tests, there is always a certain probability of an error of inference. However, if the derivational assumptions of a test are met, the probability of error is accurately known; but when they are not, the probability of error is uncertain. The two major errors of inference are: (a) concluding that population differences exist when they in fact do not; and, (b) concluding that no population differences exist when in fact they do. These two errors are referred to respectively as Type I and Type II errors, their attendant probabilities being denoted by α and β. Power is the probability of correctly concluding that population

differences exist, i.e., power = 1 - β. A statistical test is said to be robust to violation of its derivational assumptions if such violation does not alter the probability of correct or erroneous inferences based on the test, i.e., if these probabilities remain accurately known. Violation of the normality assumption, alone, does not seriously affect the accuracy of the t- and F-tests and they are therefore said to be robust to non-normality (see Glass, Peckham, and Sanders, 1972). However, violation of the homogeneity of variance assumption may have considerable effects on the t- and F-test's accuracy, especially if the groups of observations sampled from the various populations differ in size (see Glass et al., 1972). The tests are thus non-robust to variance heterogeneity. Independence of errors, the final assumption, will not be considered here, although it should be noted that the t- and F-tests are not robust to non-independence of errors (see Glass et al., 1972).

Since variance heterogeneity may substantially affect the accuracy of the t- and F-tests, alternative methods have been developed for use in this situation (e.g., Welch, 1938, 1951). In addition to these alternative statistical procedures, it is often possible to correct variance heterogeneity by transforming the original data, in which case the t- or F-tests may then be used on the transformed data. Empirical sampling studies, using computer simulated populations, have been used to compare the accuracy of the ANOVA F-test and t-test with the accuracy of such alternative statistical procedures as the Welch tests (Welch, 1938, 1951), under conditions of variance heterogeneity

and unequal group size (Brown and Forsythe, 1974a; Kohr and Games, 1974). However, no empirical studies have been conducted upon the procedure of performing an F-test or t-test upon data which has been subjected to a variance stabilizing transformation. One purpose of the present research is to make an empirical comparison of the performance of: (a) the F-test; (b) alternative statistical procedures believed to be robust to variance heterogeneity; and, (c) the F-test on transformed data.

If procedures exist that are insensitive to variance heterogeneity, why are they not always used in preference to the F-test or t-test? The answer to this question is that when variances are equal, the ANOVA F-test is more likely to find differences between population means, where they exist (i.e., has greater power), than procedures such as the Welch tests (Kohr and Games, 1974). In order to make the best use of the available tests, a researcher should test for heterogeneity of variance before testing for differences in population means. This is a practice, which is not much used, because the conventional variance tests are not robust to violation of their derivational assumption of normality (Box, 1953).

Variance tests may also be used to determine a variance stabilizing transformation. If a whole sequence of transformations are performed on the data, that set of transformed data, which gives the lowest value of the variance test statistic and permits the hypothesis of variance equality between groups to be retained, may be used in an F-test or t-test. Since theoretically a data transformation which

stabilizes variance may both fail to eliminate non-normality where it exists and introduce it where it does not exist, variance tests to be used on transformed data must be robust to non-normality.

The performance of some robust variance tests under conditions of unequal sample size and/or non-normality has been investigated recently (Games, Winkler and Probert, 1972; Gartside, 1972; Brown and Forsythe, 1974b; Martin and Games, 1976) and the results indicate that only one test, the Box (1953) test, is truly robust to both non-normality and inequality of sample size. Unfortunately, the Box (1953) test is less powerful than some of the less robust (but reasonably acceptable) tests, such as the Box-Andersen Test (Box and Andersen, 1955), or the jacknife test (Miller, 1968).

If a variance test is to be used prior to an F- or t-test of mean differences, it is difficult to decide which feature, robustness or power, is more important. A non-robust variance test may lead to erroneous rejection of the homogeneity of variance assumption which would lead to E performing a possibly less powerful test of mean equality such as the Welch test. On the other hand, a not very powerful but robust variance test may fail to detect variance heterogeneity of an order which would affect a subsequent F- or t-test of mean differences: this may lead to erroneous rejection of the null hypothesis of mean equality. The only way in which this question may be answered is to perform an empirical investigation of the sequence of variance testing followed by an ANOVA F-test, if the variance test retains the homogeneity of variance assumption, or a procedure such as the Welch

test, if the homogeneity of variance assumption is rejected. Thus a further purpose of the present research is to compare the performance of the Box (1953) variance test and some more powerful but less robust variance tests (e.g., the Box-Andersen and Jacknife tests) in choosing between alternative tests of mean differences. The question to be answered is which variance test leads to the greatest robustness and power in a subsequent test of mean differences regardless of the actual test of mean differences used.

Since the effect of unequal group size on the performance of the Box-Andersen test is not known, this will be investigated here. Also the performance of these variance tests in choosing transformations will be investigated.

Notation

At this point it is convenient to define the common notation system which will be used and also certain of the statistics which will be referred to.

Let X_{ij} represent the ith observation in the jth group where $i = 1, \ldots n_j$ and $j = 1, \ldots K$. The X_{ij} are independent variates with expected value μ_j and variance σ_j^2. The analysis of variance (ANOVA) statistic, F, is used for the comparison of K group means and is given by:

$$F = \frac{\sum_j n_j (\bar{X}_{.j} - \bar{X}_{..})^2 / (K-1)}{\sum_j (n_j - 1) s_j^2 / (N-K)} \qquad (1)$$

where

$N = \sum_j n_j$ = the total number of observations

$\bar{X}_{.j} = \sum_i X_{ij} / n_j$ = the group mean

$\bar{X}_{..} = \sum_j \sum_i X_{ij} / N = \sum_j n_j \bar{X}_{.j} / N$ = the grand mean

$s_j^2 = \Sigma (X_{ij} - \bar{X}_{.j})^2 / (n_j - 1)$ = an unbiased estimate of σ_j^2.

The numerator of the F ratio is known as the mean square between groups (MS_{MB}) and the denominator is known as the mean square within groups (MS_{WG}).

In the case where K = 2 the Student t-statistic is used instead of the ANOVA F statistic to compare the two group means. t is equal to the square root of F when K = 2 and is given by:

$$t = (\bar{X}_{.j} - \bar{X}_{.j'}) \Big/ \sqrt{\frac{(n_j - 1)S_j^2 + (n_{j'} - 1)S_{j'}^2}{(n_j + n_{j'} - 2)} (1/n_j + 1/n_{j'})} . \quad (2)$$

Effect of Assumption Violations on Type I Errors

When population means are compared the computed values of t or F are subsequently compared with a critical value at a certain percentage point in the tail of their sampling distribution. The tabled values of these statistics, at given percentage points, are calculated on the basis that the derivational assumptions of the tests have been met

and that differences between population means do not exist. The chosen region of rejection is known as the level of significance of the test or α. If the calculated value of the statistic falls within the region of rejection, the means of the populations are considered as differing from each other. However, a difference between means is not the only reason why the critical value may be exceeded. The calculated value may be one of the 5% of calculated statistics which exceed the critical value when the means are identical and the assumptions have been met, in which case a Type I error has been made. It can be seen then, that if the assumptions of the test are met, the probability of a Type I error is equal to the level of significance chosen for the test (also known as the nominal level of significance).

If the assumptions of a test are violated, the sampling distribution of the test statistic is not the same as when the assumptions are met. This means that the tabled critical values are no longer appropriate. When a nominal level of significance is chosen, the tabled critical value will not cut off the correct percentage of that sampling distribution which exists when assumptions are violated. The statistical test is now inaccurate or biased. For example, suppose the nominal level of significance is set at 5% and the tabled critical value cuts off more than 5% of the actual sampling distribution; now the probability of a calculated statistic exceeding the critical value is greater than 5%. Under these circumstances a test is biased for finding differences between means and is termed liberal. On the other hand, if the tabled critical value cuts off less than 5% of the actual

sampling distribution, the probability of a calculated value exceeding the critical value is less than 5%. Now the test is biased against finding differences between means and is said to be conservative. Thus, if a test is liberal, the actual probability of a Type I error (or the actual level of significance) is greater than the nominal level of significance and if the test is conservative, the reverse is true.

If an experimenter concludes, on the basis of a statistical test at the 5% level of significance, that mean differences exist, he has a 5% chance of being wrong. Thus the statement that mean differences exist is not a statement of fact but a probability statement based on evidence from a statistical test. The experimenter knows, that five times in 100, a value exceeding the critical value may arise by chance, when no mean differences exist. For the experimenter, or anyone else, to have faith in his probability statements, they must be accurate. Unfortunately, assumption violations can lead to inaccurate probability statements, therefore it is important to know, if possible, in what direction and to what extent the various assumption violations affect the probability of Type I errors.

Glass, Peckham and Sanders (1972, p.245) reviewed the empirical and theoretical studies on the effects of violating the assumptions upon which the derivation of the F-test is based. Specifically, in the case of the effect of heterogeneous variances upon Type I errors, their review permitted the following general conclusions: (a) When group sizes are equal, the effect of variance heterogeneity on the probability of a Type I error is generally small, but if variance

heterogeneity is extreme, this is no longer so. (b) When group sizes are unequal and the larger groups are sampled from populations with larger variances, the actual probability of a Type I error is less than the nominal level of significance. (c) When group sizes are unequal and the smaller groups are sampled from populations with larger variances, the actual probability of a Type I error is greater than the nominal level of significance. These conclusions give only a qualitative picture of the effect of heterogeneity of variance and group size; however, it is desirable to know how much a given degree of heterogeneity will affect the probability of a Type I error.

For example, in the two cases outlined below, the smaller samples have larger variances but the pattern of the sample sizes and variances is different. F-tests on both sets of data will be liberal, and presumably one more so than the other, but which set will lead to the higher value of α is not clear.

	Case A	Case B
Sample Sizes	12,23,48,72	12,18,36,90
Sample Variances	14.0,14.0,6.0,6.0	15.0,15.0,5.0,5.0

Empirical studies alone are sufficient to support Glass, Peckham and Sanders' (1972) general conclusions, but in order to make more specific statements, it is necessary to turn to a theoretical analysis of the situation.

Box (1954) provides an exact mathematical method for determining the probability of Type I error of the F-test when variances and/or

sample sizes are unequal. In addition, he provides a more accessible approximate solution in which he shows that the ratio of mean squares is distributed approximately as bF(h',h), where b, the bias coefficient, is given by

$$b = 1 + \frac{1 - 1/N}{1 - 1/K} (\bar{\sigma}^{-2} / \dot{\sigma}^{2} - 1) \tag{3}$$

where $\dot{\sigma}^{2}$, the weighted mean variance is given by

$$\dot{\sigma}^{2} = \Sigma(n_j - 1)\sigma_j^{2} / \Sigma(n_j - 1) \tag{4}$$

and $\bar{\sigma}^{2}$, the unweighted mean variance is given by

$$\bar{\sigma}^{2} = \Sigma\sigma_j^{2} / K \, . \tag{5}$$

h' and h are reduced degrees of freedom (df) given by

$$h' = \left[\sum_j (N-n_j)\sigma_j^{2}\right]^{2} / \left(\sum_j n_j\sigma^{2}\right)^{2} + N \left[\sum_j (N-2n_j)\sigma_j^{4}\right] \tag{6}$$

$$h = \left[\sum_j (n_j-1)\sigma_j^{2}\right]^{2} \Big/ \left[\sum_j (n_j-1)\sigma_j^{4}\right] \, . \tag{7}$$

When group sizes are equal the weighted mean variance is equal to the unweighted mean variance and therefore b = 1. Thus for equal group sizes the F-test is not biased, in the sense that the numerator and denominator of F have the same expected value. However, h' and h can be

shown to depend on the coefficient of variation of the variances, c, which is given by

$$c^2 = \frac{\Sigma(\sigma_j^2 - \bar{\sigma}^2)^2 / (\bar{\sigma}^2)^2}{K} \quad . \tag{8}$$

If variance heterogeneity is not present and group sizes are equal, $c = 0$, and h' and h become equal to the conventional numerator and denominator degrees of freedom for the F-test. However, if variance heterogeneity is present with equal group sizes, $c > 0$, and h' and h become less than the conventional degrees of freedom. Thus, when variance heterogeneity is present, even if group sizes are equal, use of the conventional degrees of freedom for the F-test will give actual Type I error rates which exceed their nominal values. An extreme example given by Box (1954) is the case of seven groups of three observations each, where the variances are in the ratio 1:1:1:1:1:1:10 and the actual probability of a Type I error, when $\alpha = .05$, was reported as .12 .

When group sizes are also unequal, b may assume a value greater or less than 1, which is determined by the extent of skewness of the distribution of the variances, as measured by the ratio of weighted and unweighted mean variances. If the group sizes differ, but the distribution is symmetrical (e.g., variances are say 1,2 and 3 and group sizes are 3,9 and 3), the weighted and unweighted mean variances will be equal and b will again equal 1; but, if the distribution is non-symmetrical, b may assume a value greater or less than 1. h' and h again depend mainly upon the coefficient of variation of the variances.

It can be seen that if the larger groups have larger variances, the weighted mean variance will exceed the unweighted mean variance and b will be less than 1; while if the smaller groups have the larger variances, the unweighted mean variance and b will be greater than 1. Consequently, a b value less than 1 indicates a conservative test, while a b value greater than 1 indicates a liberal test. Furthermore, the extent to which b differs from 1 is an indicator of the extent to which a given test may be conservative or liberal.

In empirical simulation studies, group size differences, group variance differences and numbers of groups are variables which, it appears, are usually chosen independently of each other so as to provide a range of values of each variable, representative of usual experimental situations. Various possible combinations of values of these variables are then tested for their effect on the probability of Type I error in an F-test of mean differences. If the intent of the study is to show the range of bias that can be introduced by inequality of group sizes and variances, this type of approach may not succeed very well. Often two apparently different patterns of variance and group size heterogeneity may give rise to the same degree of bias and only a small range of bias is demonstrated (see for example, Horsnell, 1953). However, if group sizes and variances were chosen to satisfy a wide range of b values, it would be possible to give a much more comprehensive picture of the effect of variance heterogeneity combined with unequal sample sizes. For each set of observations investigated, a b value and an associated actual probability of Type I error could be

found and any experimenter referring to this data would, after calculating his/her own sample b value, be able to infer the extent to which his/her test might be liberal or conservative.

Glass et al. (1972) reviewed two articles (Norton, cited in Linquist, 1953; and Boneau, 1960) concerning the combined effects of non-normality and variance heterogeneity on the probability of a Type I error in the t- and F-tests. Boneau's (1960) study indicated, as expected, that heterogeneneous variances produced only a small degree of bias in the t-test when group sizes were equal and populations normal; however, if the additional factor of non-normality was also present the degree of bias increased. For example, when taking samples of size five from two normal populations having variances of one and four the actual probability of a Type I error was .064 at the nominal .05 level, whereas when sampling from two rectangular populations under the same conditions the actual probability of a Type I error was .071. Actual and nominal probabilities of a Type I error were shown to be equal when sampling from two rectangular populations of equal variance.

Recent results obtained by Havlicek and Petersen (1974) confirmed Boneau's (1960) observations but in addition demonstrated the effects of sample size. Increasing sample size considerably reduced the probability of a Type I error, when taking equal sized samples from two positively skewed populations with different variances. For example, when taking samples of size five from two normal populations, one having twice the variance of the other, the probability of a Type I error for a one-tailed t-test was .0560 and .0602 in the left and right tails,

respectively (at a nominal significance level of five percent) and for the positively skewed populations, the corresponding values were .0294 and .1048. When the sample size was increased to 15, the normal population values became .0586 and .0578, whereas the values became .0338 and .0848 when sampling from positively skewed populations. It should be emphasized that when the variances were equal, there is not a substantial discrepancy between nominal and actual probabilities of a Type I error for the positively skewed distribution.

It would seem that having equal group sizes affords little protection against the effects of variance heterogeneity when sampling from skewed distributions unless the group size is comparatively large. Unequal sample sizes and variances produce even greater discrepancies between actual and nominal probabilities of a Type I error than those obtained with normal populations.

Effects of Assumption Violations on Power

Power is the probability of detecting a true difference between population means and if this true difference is not detected a Type II error has been made. An experimenter may determine a priori the power of the t- and F-tests for detecting a specified difference between population means. When a difference between population means exists the sampling distribution of t or F is different from the distribution that exists when there are no mean differences, and is referred to as the non-central t or F distribution. As can be seen from their formulae, t and F will be larger when population mean differences exist, therefore

the critical value, which cuts off say the upper five percent of t or F values when no mean differences exist, will cut off a much larger percentage of the appropriate non-central t or F distribution. The power of a test is then the probability of a calculated t or F value falling beyond the critical value in their non-central distributions. Any factor which affects the probability of a Type I error might also be expected to affect the probability of a Type II error (β) or power ($1-\beta$). Thus if an assumption violation increased the probability of a Type I error beyond its nominal value, i.e., was liberal, it would be expected to increase power and vice versa.

Control of the probability of Type I errors is important so that if mean differences are said to exist we know there is a high probability that they do in fact exist. On the other hand control of the probability of Type II errors is important so that if true mean differences do exist there is a high probability that they will be detected. For this reason it is also necessary to know how assumption violations affect the probability of a Type II error or power.

Considering first the effect of heterogeneous variances on the power of the F-test, Glass et al. (1972, p.267) state "that there exists no method by which the theoretical power of the F-test can be determined when error variances are heterogeneous." Empirical power values under any conditions, may be determined by simulation techniques: however nothing will be learned concerning the effect of heterogeneous variances unless there exists a theoretical power with which to compare the empirically determined power values. Power is a function of, amongst other

things, the non-centrality parameter which is given by

$$\delta^2 = \frac{\sum_j n_j (\mu_j - \mu..)^2}{\sigma^2} \qquad (9)$$

where $\mu..$ is the grand mean of all the populations. As can be seen, the formula for δ^2 involves a single value for σ^2, the common population variance. When variances are heterogeneous, there is no single value for σ^2 : investigators have dealt with this problem by substituting the average within group variance, $\bar{\sigma}^2$ (e.g., Horsnell, 1953; Donaldson, 1968; and Lunney, 1970).

Horsnell (1953) has shown that for equal group sizes there is a close correspondence between the actual power and the "theoretical" power calculated using $\bar{\sigma}^2$. With unequal group sizes, when the larger group has the larger variance, actual power values are less than "theoretical" power values: when the larger group has the smaller variance actual power values exceed "theoretical" power values. These results for power exactly parallel the results obtained for the probability of Type I error: this could have been predicted, since any increase or decrease in α is usually accompanied by a corresponding increase or decrease in power.

Donaldson (1968) investigated the combined effect of heterogeneous variances and non-normality on the power of the F-test. The two non-normal populations that he used were the exponential and lognormal distributions : for both distributions the arithmetic mean squared was set equal to the variance, thus any differences between the means

resulted in heterogeneous variances and the greater the mean differences the greater was the degree of variance heterogeneity. For the normal distribution the mean and variance are not related to each other and thus any pairing of means and variances may be introduced into normal distributions. Therefore, in order to obtain a proper power comparison, Donaldson (1968) made the variance equal to the mean squared for the normal population also.

For the upper ranges of ϕ ($\phi = \delta / \sqrt{K}$) Donaldson (1968) found that the F-test based upon the two non-normal distributions had higher power than that based upon the normal distribution; while at lower ϕ values the tests based on the non-normal distributions were only slightly less powerful than those based on normal distributions. The points at which the power curves crossed depended on the number of groups and group size and occurred at lower ϕ values for the lognormal than it did for the exponential distribution. For example at $\alpha = .05$, with two groups of 16 observations the lognormal and normal power curves crossed at $\phi \simeq .50$ and a power of .10 while the exponential and normal curves crossed at $\phi \simeq 1.30$ and a power of .40. The corresponding crossover points for four groups of 16 observations were at $\phi \simeq 1.10$, power = .40 and $\phi \simeq 1.25$, power = .50, respectively.

Donaldson's (1968) power curves for the normal distribution with heterogeneous variances were practically identical to those obtained when variances were homogeneous. Thus the particular pattern of variance heterogeneity seen with lognormal and exponential distributions actually provides a power advantage over the conventional situation of normal distributions and homogeneous variances, in the upper power

ranges. If an experimenter calculated power a priori, assuming equal variances and normality, it is highly unlikely that he would be satisfied with a power value below .50. Since it is only below .50 that the power of the non-normal populations becomes less than that of the normal population this would not seem to be a matter for concern, given that E determines power a priori to be > .50.

In spite of the extreme variance heterogeneity that may occur with lognormal and exponential distributions, when mean differences exist, the power of the F-test does not seem to be much affected. The explanation for this lies in the fact that when non-normal distributions are used, the numerator and denominator of the F-ratio are no longer independent of each other as they are with normal distributions. Donaldson (1968) obtained empirical correlations between the numerator and denominator of F for all the conditions which he used; it was then possible to show that "The size of the correlation co-efficient is closely associated with the degree to which F is conservative." (That is, conservative with respect to Type II errors.)

Unfortunately Donaldson (1968) did not investigate the effect of unequal group size on power in non-normal populations with heterogeneous variances. It is to be expected from what has been discussed previously that this would affect the power of the F-test considerably.

Violation of the normality assumption by itself does not usually cause major discrepancies between nominal and actual probabilities of Type I error or power. Generally, leptokurtosis increases and platykurtosis decreases power values (Glass et al., 1972). Because of the

relationship between α levels and power one might expect an increased probability of rejecting the null hypothesis when it is true (i.e., increased probability of a Type I error) for leptokurtic populations. Interestingly though, for Donaldson's two non-normal leptokurtic populations, there is the combined advantage of being both more likely to reject the null hypothesis when it is false and retain it when it is true. Donaldson attributes this doubly advantageous feature of the F-test, on his non-normal populations, to the correlation between numerator and denominator of F. Furthermore, he demonstrated that under the null hypothesis, this correlation is a function of the kurtosis of the population. Donaldson did not investigate the additional affect of unequal sample size.

In the case where variance heterogeneity is combined with unequal sample sizes the distortions of α levels and power for the t- and F-tests may become so great as to render these tests useless. The question remains as to how one may accurately test the significance of the difference between two or more means under these conditions. There are a variety of approaches to this problem and each will be considered in turn.

Alternative Procedures for Comparing Means

When there are two means to be compared and the ratio of the population variances is unknown, the problem of testing the significance of the differences between these two means is known as the Behrens-Fisher problem (Behrens, 1929; Fisher, 1935). Both the original

Behrens-Fisher solution and the later Welch-Aspin solution (Welch, 1947; Aspin, 1948) require special tables for the critical values of their respective distributions. Each of the distributions is defined by the degrees of freedom for each sample (f_j) and a quantity which is dependent on the ratio $s_j^2 / s_{j'}^2$, where s_j^2 and $s_{j'}^2$ are unbiased estimates of the population variances. Tables of the critical values of the Behrens-Fisher distribution are entered with f_j , $f_{j'}$ and $\tilde{\Theta}$, where

$$\tilde{\Theta} = \arctan \sqrt{(\lambda_j s_j^2) / (\lambda_{j'} s_{j'}^2)} \qquad (10)$$

and $\lambda_j = 1/n_j$, the reciprocal of the sample size; whereas tables of the critical values of the Welch-Aspin distribution are entered with f_j , $f_{j'}$ and

$$c = \lambda_j s_j^2 / (\lambda_j s_j^2 + \lambda_{j'} s_{j'}^2) , \qquad (11)$$

$\tilde{\Theta}$ and c are related by the following equality

$$\tilde{\Theta} = \arcsin c^{\frac{1}{2}} \text{ (see Scheffé, 1970, p.1505).} \qquad (12)$$

In addition to the Welch-Aspin asymptotic series solution, Welch (1938, 1947) has provided an approximate degrees of freedom (APDF) solution using the t-distribution. Specifically the criterion

$$v = \frac{(\bar{X}_{.j} - \bar{X}_{.j'}) - (\mu_j - \mu_{j'})}{\sqrt{(\lambda_j s_j^2 + \lambda_{j'} s_{j'}^2)}} \qquad (13)$$

(where $\bar{X}_{\cdot j}$ are sample means and μ_j are population means) follows approximately the t-distribution with degrees of freedom given by

$$f = \frac{(\lambda_j s_j^2 + \lambda_{j'} s_{j'}^2)^2}{\lambda_j^2 s_j^4 / f_j + \lambda_{j'}^2 s_{j'}^4 / f_{j'}} \qquad (14)$$

when the population means are equal.

Wang (1971) has calculated the probabilities of Type I error for each of the above tests under various conditions of population variance ratios, sample sizes and nominal levels of significance. Generally the Behrens-Fisher test was found to be rather conservative while the Welch-Aspin test showed a maximum deviation from nominal α (under the conditions investigated) of only .0009. The Welch approximate degrees of freedom test agreed very closely with the Welch-Aspin test, showing a maximum deviation from nominal α of .0018 under the same conditions. Since the Welch-Aspin critical values are available for only a selected set of α , $(f_j , f_{j'})$, and c (e.g., the smaller number of degrees of freedom must be $\geq$ 6 when α = .10, $\geq$ 8 when α = .05, and $\geq$ 10 when α = .02 or .01), it would seem more reasonable to use the Welch APDF test which only requires the t-tables. Scheffé (1970, p.1505) has stated "I judge Wang's work will justify the conclusion that Welch's approximate t-solution, . . . , is a satisfactory practical solution of the Behren's-Fisher problem."

Welch (1951) has shown that his APDF solution to the Behren's Fisher problem may be generalized to the case where there are more than

two group means to be compared. In this case the distribution of the statistic

$$W = \frac{\sum_j w_j (\bar{X}_{.j} - \tilde{X}_{..})^2 / (K-1)}{1 + \frac{2(K-2)}{(K^2-1)} \sum (1 - w_j / \sum w_j)^2 / (n_j - 1)} \qquad (15)$$

(where $w_j = 1/\lambda_j s_j'^2$, and $\tilde{X}_{..} = (\sum w_j \tilde{X}_{.j}) / (\sum w_j)$, when the population means are equal, follows approximately the distribution of F with denominator degrees of freedom given by

$$f_2 = \left[\frac{3}{(K^2-1)} \sum (1 - w_j / \sum w_j)^2 / (n_j - 1) \right]^{-1} \qquad (16)$$

and the usual numerator degrees of freedom. For large samples the numerator of W is distributed as a chi-square variable with (K-1) degrees of freedom: however this is not true for small samples. James (1951) has shown that the distribution of this quantity in small samples may be approximated by

$$\chi^2 \left[1 + \frac{3\chi^2 + (K+1)}{2(K^2-1)} \sum (1 - w_j / \sum w_j)^2 / f_j \right]$$

where χ^2 is a chi square variable with (K-1) degrees of freedom.

In an empirical investigation, Brown and Forsythe (1974a) have compared the performance of the usual ANOVA F-test, the generalized Welch APDF solution, James' solution and another solution, in which the

denominator of F is altered to have an expected value equal to the numerator when the means are equal, regardless of the value of the variances. In this latter solution the value

$$F^* = \frac{\sum_j n_j (\bar{X}_{.j} - \bar{X}_{..})^2}{\sum_j (1 - n_j / N) s_j^2} \qquad (18)$$

is distributed approximately as F with the usual numerator degrees of freedom and denominator degrees of freedom, f, given by the Satterthwaite (1941) approximation.

$$\frac{1}{f} = \sum_j c_j^2 / (n-1) , \qquad (19)$$

where $c_j = (1 - n_j / N) s_j^2 \Big/ \left[\sum (1 - n_j / N) s_j^2 \right]$.

The results of the study demonstrate the usual lack of robustness of the ANOVA under conditions of heterogeneous variances and unequal sample sizes. Of the alternatives, James' procedure gives actual probabilities of a Type I error which are greater than the nominal level of significance when the sample sizes are small (i.e., 4), while the Welch procedure and F* show reasonably good control of the probability of a Type I error. On occasions F* performs better than the Welch procedure and on other occasions the situation is reversed, however no consistent trends emerge. The Welch actual probabilities of a Type I error vary less over the conditions investigated than they do for F*, but the difference is slight. Empirical power determinations showed that F* and

the Welch procedure produced very similar results to the ANOVA when variances were equal. When variances were unequal, only F* and the Welch procedure were compared since Brown and Forsythe (1974a) considered the actual probability of a Type I error for the ANOVA to be unacceptable under these conditions. Which of the two procedures has the greater power depends upon whether the extreme means have comparatively large or small variances. Since, in the Welch procedure, means are weighted by n_j/s_j^2 and in F* by n_j, an extreme mean with a small variance would tend to increase W more than F* and conversely for extreme means with large variances. This feature makes a very sizeable difference to the empirical power of the two tests.

Kohr and Games (1974) have also compared the performance of the Welch procedure and the ANOVA under conditions of equal and unequal variances and sample sizes. Also included in their empirical investigation of Type I error rates and power were the unweighted means analysis and a procedure due to Box (1954). The unweighted means analysis employs the same MS_{WG} as the ANOVA, however, the MS_{BG} is calculated giving equal weight to each group mean and is given by

$$MS_{BG} = \tilde{n}\Sigma (\bar{X}_{.j} - \bar{G}) / (K-1) \qquad (20)$$

where $\tilde{n} = K/\Sigma (1/n_j)$

and $\bar{G} = \Sigma\bar{X}_{.j} / K$.

In the Box (1954) procedure the usual mean square ratio obtained in an ANOVA is divided by the Box bias coefficient b, calculated from the

sample data, and this statistic is referred to the F distribution with degrees of freedom h' and h (see (6) and (7) above). This study was unusual in that the extent of variance heterogeneity was quantified by calculation of the coefficient of variation of the population variances: also b values were calculated for all conditions of unequal sample size. The results show that the unweighted means analysis was even less robust than the conventional ANOVA, while the Welch procedure showed the best control of the probability of a Type I error, with the Box procedure a close second.

Results for power indicate that the ANOVA had superior power when the assumptions were met. When group sizes were equal and variances were unequal, the Welch procedure showed superior power, except when the deviant means were paired with larger variances, in which case it was less powerful than the Box procedure, which in turn was less powerful than the ANOVA. This was in accordance with Brown and Forsythe's (1974a) findings. The Box procedure was never the most powerful in the equal n case. For unequal n's and variances the Welch procedure was again usually the most powerful and when deviant means were paired with large variances the Welch procedure was again displaced as the most powerful test. The only situation in which the Box procedure was most powerful was if deviant means were all paired with large variances; it was not sufficient for one deviant mean to be paired with a large variance if another deviant mean was not. Considering the extreme specificity of the situation in which the Box procedure is most powerful, the small likelihood of knowing the situation a priori

and the small extent to which the power of the Box procedure exceeds that of the Welch procedure in these situations, it would seem that the Box procedure is not a particularly useful alternative.

All the foregoing tests belong to a general class of solutions to the problem of testing for mean differences in the presence of variance heterogeneity. Each of the tests allows for the existence of variance heterogeneity and therefore does not require the assumption of equal group variances. An alternative approach is to bend the data to fit the assumptions of the conventional ANOVA F-test or the t-test.

Removal of Assumption Violations by Data Transformation

Data transformations have frequently been used to make the data fit the assumptions of the ANOVA F-test; this procedure, although it sounds simple, is not without attendant difficulties. Firstly, the null hypothesis of mean equality is usually phrased in terms of the original data, whereas the actual hypothesis tested is on the transformed data. However, if the transformed variate has some theoretical meaning (e.g., reaction times are often transformed before analysis by a reciprocal transformation: the transformed variate then has some meaning as "speed of response") the experimenter may be perfectly willing to perform a hypothesis test on the transformed data and restrict his conclusions to the transformed variate. Problems arise when the transformation is performed merely to alter the form of the distribution of the dependent variable, in order to facilitate data analysis, and has no

value in terms of the scientific theory being tested. Under these latter circumstances the experimenter "ends up testing a different hypothesis than originally intended, and is typically not logically justified in extending his conclusions back to the original measure of interest!" (Games and Lucas, 1966, p.315). As an example, suppose the group means of the variable X are identical but the treatment populations differ from each other in some other respect such as variance. Under these circumstances an experimenter may wish to perform a variance stabilizing transformation. It is now perfectly possible that the treatment group means will differ from each other on the transformed variate, $Y = f(X)$. Thus, "rejecting the hypothesis of equal treatment means on $f(X)$ may occur because the treatment populations on X differ in variance, or in skewness, or in kurtosis, even though the population means on X are equal. Thus rejecting the hypothesis that $\mu_{f(X_1)} = \mu_{f(X_2)} = \ldots = \mu_{f(X_3)}$ implies some difference in the treatment effects, but does not clearly imply the rejection of $\mu_{X_1} = \mu_{X_2} = \ldots = \mu_{X_3}$" (Games and Lucas, 1966, p.315).

When heterogeneity of variance (heteroscedasticity) exists a transformation can be found which will stabilize variance across groups if a functional relationship exists between the mean and the variance. There are many naturally occurring forms of variation such as Poisson, binomial and lognormal distributions where the appropriate variance stabilizing transformation is known (Bartlett, 1947), however the experimenter often may not know what the shape of the population distribution is and can only work on the sample evidence. Olds, Mattson and Odeh

(1956) suggest the procedure of using sample means and variances to obtain an estimate of the regression of the population variance on the population mean. Once a regression function has been determined it is then possible to determine the variance stabilizing transformation by the method given by Bartlett (1947). Suppose the estimated relationship between the population variances and means is represented by

$$\sigma_X^2 = f(\mu_X) \tag{21}$$

where σ_X^2 is the variance on the original scale of measurements X with the mean of X equal to μ_X. Then for any transformation, which may be represented by the function g(X), the variance of this function is given approximately by

$$\sigma_g^2 = (dg/d\mu_X)^2 \; f(\mu_X) \; . \tag{22}$$

The purpose of the transformation is that the variance of the transformed variate σ_g^2 should be a constant, say c^2, thus

$$dg/d\mu_X = c/\sqrt{f(\mu_X)} \tag{23}$$

and

$$g(\mu_X) = \int \frac{c \; d\mu_X}{\sqrt{f(\mu_X)}} \; . \tag{24}$$

This integral may then be evaluated for any function of the mean (for example, when mean and variance are proportional as in the Poisson distribution, integration gives the square root transformation). Mueller (1949, p.209) notes that this procedure involves "several severe approximations," however Olds et al. (1956, p.12) maintain that "it seems to be the best one generally available." Since a transformation arrived at in this manner is only an approximation it would seem advisable to check that the variance of the transformed variate has been stabilized before proceeding with the analysis.

Transformations may be applied to the data for reasons other than achieving homoscedasticity: for example, a normalizing transformation may be required or one that removes non-additivity. It is often true that a transformation applied for one of the above reasons will incidentally achieve the remaining objectives also: but there is no guarantee that this will be true. In the method outlined above for determining the (scedasticity) transformation, it was seen that the functional relationship between mean and variance determined the transformation used; but a variety of different distribution forms may have the same relationship between mean and variance: thus the transformation which gives homoscedasticity cannot be expected to produce normality in every one of these cases (see Mueller, 1949, and Curtiss, 1943). Tarter and Kowalski (1972) have defined the precise situation in which the scedasticity transformation will also produce normality, but this will not be discussed here.

As was discussed previously, non-normality alone does not

seriously influence the conclusions drawn from the results of a t- or F-test, so it would seem relatively pointless to perform a transformation for this reason alone. In a computer simulation study, Games and Lucas (1966) have shown that performing a normalizing transformation actually caused greater deviations from theoretical power and significance levels than using the non-normal, untransformed data. Add to this the fact that the population form was known and therefore also the correct normalizing transformation, which is not the case for the typical experimenter, and the whole procedure emerges as having limited usefulness.

Another approach to achieving homoscedasticity is to choose a transformation, within a restricted family, to minimize some measure of variance heterogeneity. Box and Cox (1964) used a power family of transformations where the original variable, X, is transformed into variable Y, which is some function of X, by the equations given below.

$$Y = \begin{cases} (X^d - 1) / d & (d \neq 0) \\ \log_e X & (d = 0) \end{cases} . \qquad (25)$$

Basically the original variable X is raised to some power d, which is a curvilinear transformation and will therefore influence the subsequent ANOVA F-test. When $d \neq 0$, the remaining procedures of subtracting 1 and then dividing by d are linear transformations which have no further effect on the ANOVA F-test. This means that (25) is exactly equivalent to

$$Y = \begin{cases} X^d & (d \neq 0) \\ \log_e X & (d=0) \end{cases} \qquad (26)$$

The form (25) is preferred because it is a continuous function at d=0, since it may be shown that the limit of $Y = (X^d - 1) / d$, as d approaches zero, is $\log_e X$ (Schlesselman, 1973). This power family gives rise to many of the commonly used transformations; for example, from (26):

d = -1 gives $Y = X^{-1} = 1/X$, the reciprocal transformation,

d = ½ gives $Y = X^{\frac{1}{2}} = \sqrt{X}$, the square root transformation,

d = 0 gives Y = log X, the logarithmic transformation.

Having decided upon this family of transformations, Box and Cox (1964) attempted to arrive at a value for d (i.e., chose a transformation from within the family) which would best enable the transformed data to satisfy not only a homoscedastic model but also one which was additive and normal. The mathematical sophistication of the procedures by which the value for d was arrived at, place them beyond the scope of the present discussion.

While a procedure which attempts to satisfy all the objectives of a transformation simultaneously is theoretically appealing, it may not be particularly useful in practice as one of the objectives may be more compelling than the others. For example, in a one-way ANOVA, it is clear that there is more reason to choose d to achieve homoscedasticity than normality. Draper and Hunter (1969) have suggested that a transformation may be chosen by plotting against d, functions which occur naturally in the usual analysis. They include in these functions the

mean square (MS) ratios (F values) for treatments and interactions and a statistic which supplies information on variance heterogeneity. Thus one could choose a transformation that maximized the MS ratios for treatments, or one that minimized the MS ratio for interactions (if an additive model were preferable), or one that minimized variance heterogeneity. When choosing a transformation on the basis of maximizing the mean square ratio for a treatment, it must be clearly remembered that rejecting the hypothesis of equal group means on the transformed data may not imply its rejection on the original data. Under these circumstances, making inferences in terms of the original untransformed variable may be very tempting, but it is dangerous. Maximizing mean square ratios to obtain significant treatment effects, or minimizing them to simplify the theoretical model are not, per se, sufficient reasons for choosing a particular transformation, since variance heterogeneity can bias the F-test and reducing it should be a primary not secondary goal of the transformation. If a transformation, which achieves homoscedasticity, also maximizes F values for treatment effects, so much the better. Also, if the transformed variable is meaningful, the transformation is even more valuable since this will lessen the temptation to make inferences in terms of the original variable.

When choosing a transformation to achieve homoscedasticity some measure of the attainment of this objective is necessary. Testing for variance homogeneity prior to an ANOVA test for equality of means has traditionally been regarded as pointless because of the notorious sensitivity of variance tests to non-normality. As Box (1953, p.333) states

"To make the preliminary test on variances is rather like putting out to sea in a rowing boat to find out if conditions are sufficiently calm for an ocean liner to leave port!" This problem is of great concern in testing the efficiency of a transformation since it is possible for the transformation to achieve homoscedasticity without achieving normality. Fortunately tests for the equality of several ($K \geq 2$) variances which are robust to non-normality are now available, however they seem to have lower power than the less robust alternatives.

Procedures for Comparing Variances

Bartlett's test (Bartlett, 1937) has traditionally been used as a test of variance homogeneity: the test statistic is given by:

$$M = (N-K) \log_e MS_{WG} - \sum_j (n_j - 1) \log_e s_j^2 \quad . \qquad (27)$$

It may be shown that when the null hypothesis of equal group variances is true and provided the parent population is normal, M is distributed in large samples as χ^2 with K-1 degrees of freedom, while for small samples the quantity M/(1+A) has approximately the same distribution (Bartlett, 1937). A is an adjustable constant which tends to zero for large group sizes and is given by

$$A = \left[1/3(K-1)\right] \left[\sum_j (1 \; / \; (n_j - 1) - 1 \; / \; (N-K) \right] \quad . \qquad (28)$$

Box (1953) has shown that M is distributed asymptotically not as χ_{K-1}^2 but as $(1 + \gamma_2 \; / \; 2) \; \chi_{K-1}^2$ where γ_2 is a measure of the kurtosis of the parent population given by

$$\gamma_2 = \sum_j \left[(X_{ij} - \mu_j)^4 / \sigma_j^4 \right] - 3 \quad . \qquad (29)$$

In normal populations the value of γ_2 is zero while for platykurtic populations it is less than zero and in leptokurtic populations it is greater than zero. Thus, if the parent population has a different degree of kurtosis than a normal distribution, M will be asymptotically biased. In large samples the mean of the distribution curve of M would be $(1 + \gamma_2 / 2)$ (K-1) instead of K-1 and the standard deviation $(1 + \gamma_2 / 2)$ $[2(K-1)]^{\frac{1}{2}}$ instead of $[2(K-1)]^{\frac{1}{2}}$. Thus the discrepancy in means relative to the standard deviation would become larger as the number of groups, K, increased, accentuating the tendency for a liberal or conservative test in leptokurtic and platykurtic populations, respectively. In fact Box (1953) showed that the sensitivity to non-normality of Bartlett's criterion compares favourably with that of existing tests for normality!

Since the distribution of Bartlett's criterion is affected by the kurtosis of the parent population, Box and Andersen (1955) proposed a modification of the M-test which corrects M by a factor containing sample estimates of the fourth moment from the mean, which is a measure of kurtosis. Specifically Box and Anderson (1955) define $M' = M/(1+.5c_2)$ where M is Bartlett's M and $c_2 = K\Sigma k_{4j} / (\Sigma s_j^2)^2$. k_{4j} is given by $k_{4j} = n(n+1)\Sigma x^4 - 3(n-1)\ (\Sigma x^2)^2 / (n-1)(n-2)(n-3)$ where x's are deviations from the group mean. A small sampling study performed by Box and Andersen (1955) indicated that while M' was a considerable improvement upon M, it was still sensitive to non-normality particularly in platykurtic populations.

Box (1953, p.330) states that "asymptotically the M test is like an analysis of variance on the sample variances instead of sample means, but the quantity . . . corresponding to the between-groups mean square is compared not with an estimate from the internal evidence of the samples but with a theoretical value of the variance which is appropriate only when the parent distribution is normal." He then goes on to suggest that a criterion less sensitive to kurtosis may be found by utilizing the information on the variation to be expected in the sample variances, which may be gathered from the internal evidence in the samples. To this end he suggests breaking up the groups, whose variances are to be compared, into subsamples and then performing an analysis of variance upon the logarithms of the subsample variances. Bartlett and Kendall (1946) suggested the logarithmic transformation for use with variance data. Since the mean of the distribution of sample variances is proportional to its variance, this transformation would be expected to stabilize variance. The results of a small sampling study on a rectangular population presented by Box (1953) indicated the greater robustness of his suggested method compared to Bartlett's test.

Examples of other tests which do not use evidence on variance variability within samples are those proposed by Cochran (1941) and Hartley (1950). Cochran's criterion is the ratio of the largest group variance to the sum of the group variances, for which tabled values of the upper percentage points are available; while Hartley's F_{max} test refers the ratio of the largest over the smallest variance to the tables of the F distribution. Box (1953) calculated actual probabilities of

exceeding the nominal five percent level when the null hypothesis was true for F_{max}, and found similar discrepancies to those found using M when working with a non-normal parent population. He concluded also that Cochran's test might be expected to show similar deviations.

Games, Winkler and Probert (1972) performed an empirical sampling study in which they confirmed Box's (1953) findings concerning the lack of robustness of the Bartlett, Cochran and Hartley tests to non-normality while at the same time showing the excellent robustness of Box's suggested procedure of performing an analysis of variance on the logarithms of the subsample variances. The power of Box's procedure is considerably less than that of Bartlett's test especially if the smallest possible subsample size of two is used. If a large number of subsamples is used this gives a greater number of degrees of freedom for MS_{WG} and thus greater power, however the larger the number of subsamples the smaller the subsample size which leads to a greater expected value of MS_{WG} and thus lower power. Since subsample number and size are inversely related to each other and affect power oppositely, there must clearly be an optimum value for subsample size given the sample size. Games et al. (1972) came to the conclusion that for sample sizes from 12 to 18 a subsample size of three yielded optimum power with very little loss of power up to sample sizes of 36, while for the larger samples (i.e., > 36) it made very little difference whether the subsample sizes were four, five, or six.

Games et al. (1972) also demonstrated that the Box-Anderson procedure was less robust than the Box (1953) procedure discussed in the

preceding paragraph: it produced a liberal test in all but a symmetric leptokurtic population, where the test was almost exact, and a rectangular population where the test was extremely conservative. In normal populations the power of the Box-Andersen test exceeds that of the Box (1953) procedure but in Games et al.'s (1972) extremely skewed and symmetric leptokurtic populations the power of the tests was practically identical if the Box subsample size was three.

Overall and Woodward (1974) have proposed a Z-variance test which has the advantage over Bartlett's test of simplicity and easy generalizability to complex factorial designs for the purpose of analyzing variance heterogeneity as a treatment effect. This test also carries the objection that it does not utilize internal evidence from the treatment groups concerning variability of variance estimates, which makes it susceptible to non-normality. Levy (1975) has compared the Z-variance test and the Box (1953) test under varying conditions of non-normality: he found, as predicted by Overall and Woodward (1974) that the Z-variance test was not robust to non-normality. Levy (1975) confirmed Games et al.'s (1972) findings on the robustness of the Box (1953) procedure and also found it to have very low power in comparison to the Z-variance test on normal populations. For the Box (1953) procedure Levy (1975) used a subsample size of two which was shown by Games et al. (1972) to produce a power of about half that obtained when using a subsample size of three. Thus Levy (1975) demonstrated the power of the Box (1953) procedure under the most unfavourable conditions.

Levene (1960) proposed two forms of a test of variance homogeneity which he found relatively robust to non-normality. One form of the test

is an analysis of variance upon the absolute deviations of observations from their group mean, while the second form uses squared deviations in place of the absolute deviations. Miller (1968) has shown that the absolute deviation form is not asymptotically distribution free and should therefore not be robust to non-normality: Levene (1960) did in fact observe this in his sampling study as did Games et al. (1972). All of the above authors found the squared deviation form to be relatively robust. In comparing the power of the alternative forms of the test, both Levene (1975) and Games et al. (1972) found the power of the absolute deviation form to be greater.

Levene (1960) also stated that the power of the absolute deviation form of his test was comparable to that of the Box-Andersen test, one of the more powerful tests available. Miller (1968) compared the power of the squared deviation form to that of the Box-Andersen test and found it to be slightly less in samples of size 25 but considerably less in samples of size 10, while Games et al. (1972) found an even lower power in samples of size six. Thus the relative inferiority of the squared deviation form of Levene's test increases as sample size decreases. Games et al. (1972) attributed this phenomenon to the fact that the squared deviation values are not independent of each other and the degree of dependence increases as sample size decreases: the same is also true for absolute deviations.

Brown and Forsythe (1974b, p.366) have proposed an adaptation of the absolute deviation form of Levene's (1960) test which makes it more robust to non-normality. They recommend that "when departures from

normality are anticipated, the estimate of the mean for each group in the Levene statistic should be replaced by a more robust estimate of central location." Thus the 10 percent trimmed mean (the mean obtained after deleting the 10 percent largest and the 10 percent smallest values in a group) is recommended for long-tailed distributions and the median for skewed distributions. Use of the median brought the actual probabilities of Type I error very close to their nominal values in both a long-tailed distribution (Student's t on 4 df) and a skewed distribution (Chi square on 4 df); however, use of the 10 percent trimmed mean was only effective for the long-tailed distribution but did not make the test robust with the skewed distribution. Substitution of the median for the mean produced only slight power losses with either a normal, a long-tailed or a skewed distribution, provided the sample size was large ($n_j = 40$); however, with small samples ($n_j = 10$) a dramatic power loss resulted. Fellers (cited in Martin and Games, 1976) has shown that with even smaller samples ($n_j = 5$) the ANOVA on absolute deviations from the median produces erratic, uninterpretable results.

Miller (1968) applied the jacknife procedure to testing hypotheses on variances in the two group case and the procedure was subsequently generalized to the $K > 2$ group situation by Layard (1973). In this procedure, the observations in each group are divided into p_j subgroups and variance estimates are made on the remaining observations in each group after deleting the ℓ^{th} subgroup: each of the p_j subgroups is deleted in turn thus giving a total of p_j variance estimates ($s_{j-\ell}^2$)

in each group. For each of the $s_{j-\ell}^2$'s a new estimate, $\theta_{j\ell}$, is formed:

$$\theta_{j\ell} = p_j \log_e s_j^2 - (p_j - 1) \log_e s_{j-\ell}^2 \qquad (30)$$

The jacknife test statistic is an F statistic from a one-way ANOVA on the $\theta_{j\ell}$, namely

$$J = \frac{\sum_j p_j (\theta_{j\cdot} - \theta_{\cdot\cdot})^2 / (K-1)}{\sum\sum_{j\ell} (\theta_{j\ell} - \theta_{j\cdot})^2 / \sum_j (p_j - 1)} , \qquad (31)$$

where $\theta_{j\cdot} = \sum_\ell \theta_{j\ell} / p_j$

$$\theta_{\cdot\cdot} = \sum\sum_{j\ell} \theta_{j\ell} / \sum_j p_j .$$

In Miller's (1968) sampling study the actual probability of a Type I error and the power of the jacknife test were both shown to be approximately equivalent to the values obtained for the Box-Andersen test, provided the subsample size for the jacknife test was one.

Since unequal group sizes are a common occurrence, it is important that tests of homogeneity of variance should be robust to inequality of group size. If a test of the homogeneity of variance assumption is performed prior to an ANOVA F-test of mean equality, it is even more important that the variance test should not be biased by unequal group size. This is firstly, because the F-test on means is not itself affected by inequality of group size per se, secondly, because the effect of variance heterogeneity on the F-test of means is most

pronounced when group sizes also differ and thirdly, because a variance test permissively biased by unequal group sizes, could erroneously reject the homogeneity of variance assumption and lead E to perform a less powerful test of mean equality, such as the Welch procedure, which is robust to heterogeneous group variances.

Brown and Forsythe (1974b) performed an investigation of the effect of unequal group sizes and non-normality on the robustness and power of certain variance tests, including the jacknife and Levene tests, in the two group situation. They noted that the size of the jacknife statistic was larger than it should be when group sizes differed and suggested that this was probably due to the lack of robustness of the ANOVA when the within group variances were unequal. In the jacknife procedure the variance estimates, $s_{j-\ell}^2$, are calculated from a larger number of observations in larger groups and they will therefore be more stable in larger groups. Thus the variance of the variance estimates within each group will be less for larger groups and more for smaller groups. This pairing of smaller within group variances and larger group sizes is known to produce a liberal bias in the ANOVA F-test of group mean differences hence the jacknife procedure should always be liberal in the presence of unequal group sizes. Martin and Games (1976) confirmed Brown and Forsythe's (1974b) findings of a liberal bias in the jacknife test, when n_j's are unequal, for the more general condition of $K > 2$ groups ($K = 3$ in this case). It should be noted that Brown and Forsythe (1974b) did not detect a permissive bias in the jacknife test with unequal n_j's and normal distributions and Martin and Games (1976)

found significant differences between actual and nominal probabilities of a Type I error in only one out of four sets of 1,000 simulated analyses in this condition. Apparently, unequal n_j's, per se, do not have much effect on the probability of a Type I error for the jacknife test in spite of the rationale outlined above. It is only in combination with non-normality that unequal n_j's have an effect: the already permissive bias created by non-normality is considerably augmented by the introduction of unequal n_j's.

Brown and Forsythe (1974b) demonstrated that their modification of the Levene test using absolute deviations from the median was robust to unequal group size in both normal and non-normal populations. However, they only investigated the two group situation for two conditions of unequal group size, namely, $n_j = 10$, $n_{j'} = 20$ and $n_j = 20$, $n_{j'} = 40$.

Martin and Games (1976) also investigated the effect of non-normality and unequal group size on the probability of a Type I error and power of three forms of the Box (1953) procedure. In addition to the original Box (1953) procedure, they investigated Scheffé's (1959, p.83) modification of the test and a modification due to Bargman which was introduced by Gartside (1972). The original Box (1953) test has the form

$$\frac{\sum_j p_j \, (Y_{j.} - Y_{..})^2 \, / \, (K-1)}{\sum\sum_{j\ell} (Y_{j\ell} - Y_{j.}) \, / \, \sum_j (p_j - 1)} \quad , \tag{32}$$

where $Y_{j\ell} = \log_e s_{j\ell}^2$,

$Y_{j\cdot} = \Sigma_{\ell} Y_{j\ell} / p_j$,

$Y_{\cdot\cdot} = \Sigma\Sigma_{j\ell} Y_{j\ell} / \Sigma_j p_j$,

$s_{j\ell}^2$ = variance estimate calculated on the ℓ^{th} subgroup in the j^{th} group,

which is an ANOVA on the logarithms of the subgroup variances. The Scheffé (1959) adaptation was designed to accommodate unequal subgroup sizes (as might occur when the subgroup size is not a factor of the group size) and sampling from non-normal populations: it has the form

$$\frac{\Sigma_j v_j (\eta_{j\cdot} - \eta_{\cdot\cdot})^2 / (K-1)}{\Sigma\Sigma_{j\ell} v_{j\ell} (Y_{j\ell} - \eta_{j\cdot})^2 / \Sigma_j (p_j - 1)} \quad , \tag{33}$$

where $v_{j\ell}$ = degrees of freedom upon which $s_{j\ell}^2$ is based,

$v_j = \Sigma_{\ell} v_{j\ell}$

$\eta_{j\cdot} = \Sigma_{\ell} v_{j\ell} Y_{j\ell} / v_j$

$\eta_{\cdot\cdot} = \Sigma\Sigma_{j\ell} v_{j\ell} Y_{j\ell} / \Sigma\Sigma_{j\ell} v_{j\ell}$.

As can be seen this procedure weights the contribution of each subgroup variance estimate according to its degrees of freedom. The Bargmann modification (Gartside, 1972) was designed to accommodate unequal group size and uses two constants one of which is added to the $\log_e s_{j\ell}^2$ value

and another which weights this combined value. Thus the variable $Y_{j\ell} = \log_e s_{j\ell}^2$ is replaced by the variable $z_{j\ell} = w_{j\ell}(\log_e s_{j\ell}^2 + c_{j\ell})$ where $c_{j\ell}$ is given by

$$c_{j\ell} = 1/v_{j\ell} + 1 / (3v_{j\ell}^2) \quad , \tag{34}$$

and the weighting constant $w_{j\ell}$ is given by

$$1/w_{j\ell} = 2/v_{j\ell} + 2/(v_{j\ell}^2) + 4/(3v_{j\ell}^3) \; . \tag{35}$$

These constants are used to remove bias and to satisfy better the homoscedasticity assumption of the ANOVA when the n_j's are unequal. The test statistic is:

$$\frac{\sum_j w_j \, (\eta'_{j\cdot} - \eta'_{\cdot\cdot})^2 \; / \; (K-1)}{\sum\sum_{j\ell} w_{j\ell} \, (z_{j\ell} - \eta'_{j\cdot})^2 \; / \; \sum_j (p_j - 1)} \tag{36}$$

where $\eta'_{j\cdot} = \sum_\ell w_{j\ell} \, z_{j\ell} \; / \; w_j$

$\eta'_{\cdot\cdot} = \sum\sum_{j\ell} w_{j\ell} z_{j\ell} \; / \; \sum\sum_{j\ell} w_{j\ell}$

$w_j = \sum_\ell w_{j\ell}$.

The condition where group sizes are equal but subgroup sizes are not must always occur if n_j is a prime number: e.g., if $n_j = 7$ then the group of observations may either be divided into three groups of size two, two and three or two groups of size three and four, bearing in mind that a variance estimate can only be calculated on two or more

observations. Under this type of condition (equal n_j of 7 or 17), Martin and Games (1976) found both the Box and Scheffé procedures to be robust to non-normality while the Bargman modification was robust under moderate population leptokurtosis ($\gamma_2 = 3.0903$) but not under extreme leptokurtosis ($\gamma_2 = 6.0041$). The Scheffé procedure was more powerful than the Box procedure and the Bargman procedure more powerful than the Scheffé.

When group sizes are unequal the subgroup size may be maintained constant across groups or may be increased with increase in group size. In the situation where each different group size may be factored by a constant subgroup size the Box, Scheffé and Bargman procedures are identical: Martin and Games (1976) found the procedure robust with three groups of size 6, 12 and 18 and a constant subgroup size of three in normal and non-normal populations. When each different group size cannot be factored by a constant subgroup size, the situation is similar to that in the preceding paragraph except that the group sizes are now different. Martin and Games, (1976) found essentially the same results in this situation as with equal group sizes except that the Bargman procedure was now not robust in the moderately leptokurtic population either. Increasing subgroup size with group size caused liberal Type I error rates for both the Box and Scheffé procedures but the Bargman procedure was robust even in the extremely leptokurtic population, unfortunately however, at the expense of power.

When group sizes are unequal, the Box and Scheffé procedures with constant subgroup size across groups and the Bargman procedure

with subgroup size increasing as group size increases are all robust. However, of the three, the Box and Scheffé procedures are to be preferred, as they produce greater power on the same sets of data. The Scheffé procedure is slightly more powerful than the original Box procedure under all conditions where both are robust.

On comparison of all of the available tests of variance homogeneity, it becomes apparent that the most robust tests are the Box (1953) procedure, its modification by Scheffé (1959), and Brown and Forsythe's adaptation of the Levene test. On the other hand, the jacknife and Box-Andersen tests are more powerful; and although they are not as robust as the Box and Scheffé procedures, they still perform much better than the conventional Bartlett test. Martin and Games (1976, p.13) found that if the jacknife test is used with a nominal alpha of .01, "the true risk of a Type I error is approximately .05 or less" and power "is approximately equal to that of the Box tests when the population is leptokurtic." It seems that the Box and jacknife tests are equivalent if nominal alpha is reduced for the latter. If this is indeed so (Martin and Games do not present any data on this), then the Box test is still preferable as it is easier to compute. Also the one advantage of the jacknife test, its power, is lost. From the preceding information it seems that the usual situation of paying for increased power by losing robustness and vice versa, also applies to tests on variances.

There are two points of view regarding desirable features of variance tests to be used prior to a test of mean differences. The rationale for choosing a very robust test such as the Box procedure is

that one would not wish to abandon an ANOVA F-test on means and proceed to a less powerful test of mean equality, just because a liberal variance test had caused erroneous rejection of the assumption of variance homogeneity. If one were using, say Bartlett's test on a non-normal population, this is precisely the situation that might occur. The ANOVA F-test on means is not affected by non-normality while Bartlett's test is. Thus Bartlett's test might find variance heterogeneity where none exists, and the ANOVA F-test on means might be abandoned for no reason. This is the type of situation which prompted Box (1953) to suggest that prior tests on variance were pointless. The other side of the coin is presented by Kohr and Games (1974, p.67) who subscribe to the opinion "that with small n_j's E may have such low power on his test of homogeneity of variance that he fails to detect more extreme variance conditions" (Kohr and Games, 1974, p.67). If a degree of variance heterogeneity, which would substantially affect the probability of a Type I error in an ANOVA F-test of mean equality, were not detected by a robust variance test, then again the test is pointless: the ANOVA F-test on means might now be done under conditions where its probability of a Type I error was high, and the very situation the variance test was designed to protect against might in fact occur.

However there is one feature of variance tests, evident in both Brown and Forsythe's (1974b) and Martin and Games' (1976) data, that would mitigate against the problem of insufficient power to detect a degree of variance heterogeneity which might affect a subsequent ANOVA F-test on means: the Box test especially has more power to detect

heterogeneity of variance when small group sizes are paired with large variances, which is precisely the situation which gives rise to liberal Type I error rates in an F-test of mean equality. When large group sizes and large variances are paired, the Box test is less powerful, and this is the situation which gives rise to conservative Type I error rates in an ANOVA F-test of mean differences. Thus, the variance test has more power where it is needed (from the point of view of the ANOVA F-test on means) and less power where it is not needed.

To date the performance of tests on variances and tests on means have been empirically studied only in separate investigations. If one wishes to discover how a variance test will perform in deciding between a test of means which is not robust to variance and group size heterogeneity (such as the ANOVA F-test) and one which is (such as the Welch test), it is necessary to perform both the test of variance homogeneity and the test of mean equality recommended by the variance test on the same set of data. If the variance test functions well in its capacity of choosing between tests on means, a more robust and more powerful overall test of mean differences should result. It is intended here to compare the performance of several variance tests, which cover a range of robustness (and consequently power), in making effective choices between alternative tests of mean differences.

A data transformation to remove heteroscedasticity has been suggested as a method of overcoming the problem of using the ANOVA F-test of the equality of group means in the presence of unequal group variances. Since there are problems with this approach, such as finding the appropriate transformation and the necessity of confining inferences to the

transformed variable, it would seem important to determine if such a procedure does indeed provide a viable solution, and if so, under what conditions. The ready availability of computer programmes for the ANOVA with several transformation options, and also the general familiarity of the ANOVA make data transformations an attractive solution to the variance heterogeneity problem. However, alternative statistical tests such as Welch's (1951) W and Brown and Forsythe's (1974a) F* do not suffer from the difficulties of use and interpretation that are inherent in the use of data transformations. Also, although the calculation of W or F* is more cumbersome than that of the ANOVA F, computer programmes for their calculation are easily prepared.

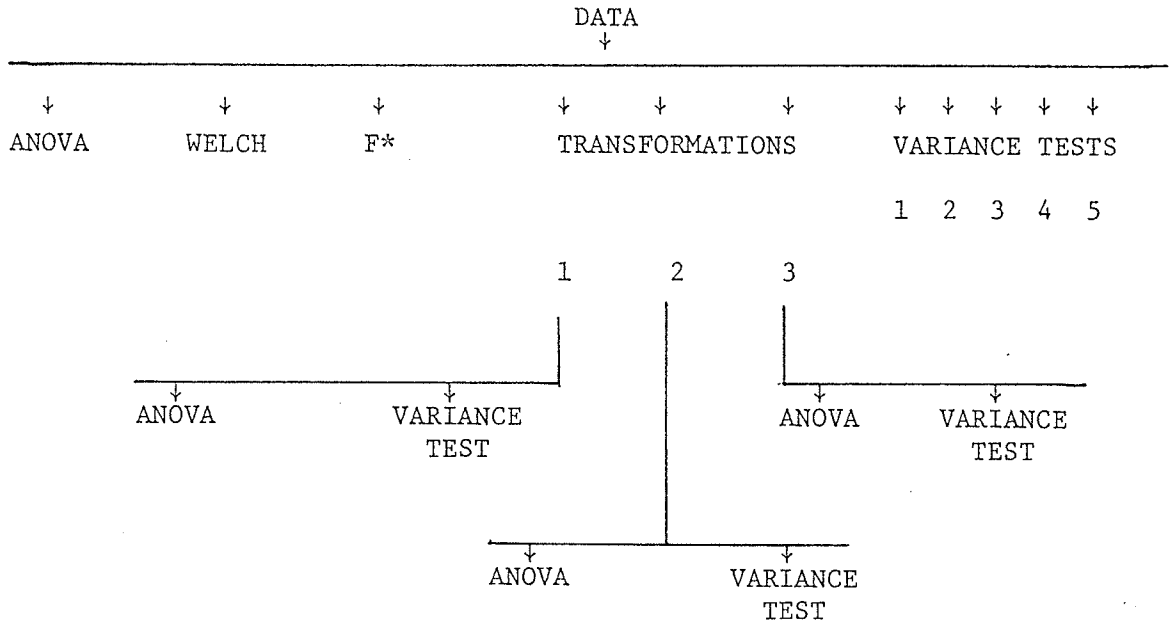
Use of data transformations to correct variance heterogeneity is a common procedure, but it is highly probable that the procedure may frequently be used incorrectly, as persons naive to the problems of data transformation may mistakenly make inferences in terms of the untransformed variable. Since investigators often apply standard variance stabilizing transformations where they are not appropriate, it is proposed that the probability of a Type I error and power, in this investigation, be determined for the ANOVA F-test on means after a variance stabilizing transformation, and that these empirically determined values be compared with the a priori determined alpha and power values. Thus, for example, if the populations sampled do not differ in their means but only in their variances, transformation and subsequent testing of mean

differences by the F-test should lead to rejection of the null hypothesis of mean equality on the original variable no greater percentage of times than that indicated by nominal alpha. However, it is entirely possible that the transformation while removing variance heterogeneity, may introduce mean differences on the transformed variable and thus the null hypothesis of mean equality on the original variable may be rejected a greater percentage of times than that indicated by nominal alpha. If this is the case the transformation creates a new problem of interpretation even though it may remove the original problem of variance heterogeneity. Thus it is important to determine if anything is to be gained by the procedure of comparing means via an ANOVA F-test on data transformed to eliminate variance heterogeneity.

In this investigation both a normal and a non-normal population will be used. The non-normal population will, like the normal population, have a mean and variance which are not functionally related. It is only rarely in behavioural research that grossly non-normal populations are encountered; more often the population is of a type which could be called a "contaminated" normal population (see Andrews, Bickel, Hampel, Huber, Rogers and Tukey, 1972, p.60). These populations may have slight or moderate skewness and kurtosis but do not have a well-defined distributional form such that a specific relationship exists between the mean and variance. Thus it would in most cases prove difficult to derive the appropriate variance stabilizing transformation by the method outlined previously (pp. 28-29).

When sampling from either population the following chart

indicates the sequence of procedures which will be performed on each set of computer generated data



Data will be generated under a variety of population and sampling conditions: the variables to be manipulated are degree of variance heterogeneity across treatment populations, degree and pattern of differences in treatment population means and group size (equal and unequal across treatment groups). The combinations of group sizes and variances will be chosen so as to satisfy a wide range of values of the Box bias coefficient, b. For each combination of treatment population and sampling conditions 2,000 sets of data will be generated and thus the above chart will be followed 2,000 times. The percentage of the 2,000 statistics for each procedure that falls beyond the critical value gives either the probability of a Type I error (when treatment effects are absent), or

power (when treatment effects are present).

As can be seen from the preceding chart the performance of the ANOVA F-test, the Welch test, F* and the ANOVA F-test following each of three data transformations may be compared directly under a variety of conditions. This comparison will answer the question of which test is the best overall for a given set of conditions. However, since E does not ever know what the population conditions are, he may wish to perform a sample variance test before proceeding to a test on means. Thus the above chart will be used to simulate the sequence of procedures an actual E might perform. The chart includes several variance tests but for now one only will be considered. Suppose Scheffé's modification of the Box procedure (Box-Scheffé test) has been performed on the data: if the Box-Scheffé test is significant, E would proceed to say a Welch test and, if it is not, E would proceed to an ANOVA F-test. It is important to know what the probability of a Type I error and power in testing mean differences are for this whole procedure. Thus, in the 2,000 simulations performed on each set of population and sampling conditions, the number of significant ANOVA F-test results occurring with insignificant Box-Scheffé test results will be added to the number of significant Welch test results occurring with significant Box-Scheffé test results. For each variance test used two overall procedures may be compared: the variance test choosing between an ANOVA F-test or a Welch test and the variance test choosing between an ANOVA F-test or F* test.

In the case of transformations the percentage of ANOVA F

statistics falling beyond the critical value will be counted in two ways. Regardless of the presence or absence of variance heterogeneity in the populations sampled it is probable that, due to sampling fluctuation, different transformations and sometimes no transformation will be recommended for each sample on the basis of a variance test. Thus for one sample no transformation, another sample a square root transformation and another a logarithmic transformation may be recommended by the variance test. Regardless of which transformation (e.g., none, square root, logarithmic or reciprocal) precedes the ANOVA F-test of mean equality, all these procedures will be considered equivalent and counted together. A second method of counting ANOVA F statistics will be for each transformation regardless of the variance test results. Thus, on the one hand, the usefulness of the procedure of choosing transformations on the basis of a variance test may be evaluated, and, on the other hand, the usefulness of a specific transformation for specific population conditions may be found.

The use of both normal and non-normal populations will allow determination of the robustness of the Welch and F* procedures to non-normality. It is probable that these procedures are as robust as the ANOVA F-test to non-normality, since both are based on the F distribution, however this does remain to be demonstrated.

When variance tests are used to detect differences in variance between treatment groups, they may often be used when treatment population means differ. If E is concerned with hypotheses on variances, it is possible that population mean differences also exist but if E is mainly concerned with hypotheses on means, and uses a variance test

merely as a test of the ANOVA homogeneity of variance assumption, it is not only possible but probable that mean differences exist. However the performance of variance tests has not so far been investigated in the presence of mean differences: the design of the present study will permit this investigation.

Method

Probability of a Type I error and power was empirically determined for several individual statistical tests and also several combinations of statistical tests under a variety of simulated population conditions and for different patterns of sampling from the simulated treatment populations. The tests and combinations of tests evaluated were:

A. Tests for variance homogeneity

   1. the Bartlett test (see p. 33)
   2. the Box-Andersen test (see p. 34)
   3. the Brown and Forsythe test (see pp. 38-39)
   4. the Miller jacknife test (see pp. 39-40)
   5. the Box-Scheffé test (see p. 43)

B. Tests for mean equality

   1. the ANOVA F-test (see pp. 5-6)
   2. the Welch test (see p. 22)
   3. Brown and Forsythe's F* test (see p. 23)
   4. the ANOVA F-test following a logarithmic transformation (T_1)
   5. the ANOVA F-test following a square root transformation (T_2)
   6. the ANOVA F-test following a reciprocal transformation (T_3)
   7. the ANOVA F-test subsequent to whichever of the following procedures gives the lowest value of the Box-Scheffé statistic: logarithmic transformation; square root transformation; reciprocal transformation; no transformation (T_4)

C. Combinations of means tests following testing for variance homogeneity

1. ANOVA F test following a non-significant Box-Scheffé test and Welch test following a significant Box-Scheffé test (FW/BS).

2. ANOVA F test following a non-significant Miller's jacknife test and Welch test following a significant Miller's jacknife test (FW/JK).

3. ANOVA F test following a non-significant Brown and Forsythe variance test and a Welch test following a significant Brown and Forsythe variance test (FW/BF).

4. ANOVA F test following a non-significant Box-Andersen test and a Welch test following a significant Box-Andersen test (FW/BA).

5. ANOVA F test following a non-significant Bartlett test and a Welch test following a significant Bartlett test (FW/B).

6. ANOVA F test following a non-significant Box-Scheffé test and a Brown and Forsythe F* test following a significant Box-Scheffé test (FF*/BS).

7. ANOVA F test following a non-significant Miller's jacknife test and a Brown and Forsythe F* test following a significant Miller's jacknife test (FF*/JK).

8. ANOVA F test following a non-significant Brown and Forsythe variance test and a Brown and Forsythe F* test following a significant Brown and Forsythe variance test (FF*/BF).

9. ANOVA F test following a non-significant Box-Andersen test and a Brown and Forsythe F* test following a significant Box-Andersen test (FF*/BA).

10. ANOVA F test following a non-significant Bartlett test and a Brown and Forsythe F* test following a significant Bartlett test (FF*/B).

When simulating the various conditions under which the performance of the above tests and test sequences were assessed, the following model for the one-way fixed effects ANOVA was used:

$$X_{ij} = \mu_j + e_{ij} \quad , \; i = 1,\ldots n_j \quad , \; j = 1,\ldots K \quad . \tag{37}$$

Each individual observation, X_{ij} , was considered to be composed of two elements: μ_j , the j^{th} population mean, and e_{ij} , the random error associated with each individual observation. The e_{ij}'s are normally distributed with a mean of zero and variance σ_e^2 . In addition to the simulation of normally distributed errors, a non-normal population distribution form was used to investigate robustness to non-normality: this was a chi square distribution with two degrees of freedom, χ_2^2, which is extremely leptokurtic.

Simulation of sampling from a normal distribution proceeded according to the method of Marsaglia, MacLaren and Bray (1964). In this method, pairs of independent pseudorandom numbers (U_1 , U_2), in the range of zero to one, are generated from a rectangular distribution and are then transformed into pairs of normally distributed pseudorandom numbers (Z_1 , Z_2) with a mean of zero and variance one N(0,1) by the relationship:

$$\begin{aligned} Z_1 &= (-2 \log_e U_1)^{\frac{1}{2}} \cos 2\pi U_2 \\ Z_2 &= (-2 \log_e U_1)^{\frac{1}{2}} \sin 2\pi U_2 \quad . \end{aligned} \tag{38}$$

Pseudorandom numbers distributed as χ_k^2 may be obtained from independent random normal deviates through the relationship:

$$\chi_k^2 = \sum_{i=1}^{k} Z_i^2 \quad . \tag{39}$$

Thus, to obtain pseudorandom numbers distributed as χ_2^2, two squared random normal deviates were summed.

Since the mean of a chi-square distribution is equal to its df and the variance is equal to 2df, it was necessary to scale the chi-square variates so that their distribution had the same mean and variance as the normal, N(0,1), distribution. Thus, the χ_2^2 distribution had a mean and variance of two and four, respectively, before scaling; and in order to give it a mean of zero and a variance of one, each χ_2^2 variate firstly had two subtracted from it and then was divided by the square root of four. This procedure resulted in a skewed, leptokurtic distribution having the same mean and variance as the normal distribution. The skewness (γ_1) and kurtosis (γ_2) of the χ_2^2 distribution are theoretically equal to two and six, respectively. Values of skewness and kurtosis for unit normal populations are both theoretically equal to zero. Both normal and skewed variates were distributed to four treatment groups.

Initially, all distributions had a variance of one, but in order to assess the power of the variance tests and robustness of the means tests, varying degrees of heteroscedasticity were simulated. Unequal variances across groups were obtained by multiplying the generated variables within each treatment group by the standard deviation required for each level of the treatment variable such that the unweighted mean variance was equal to one. The degree of variance heterogeneity was indexed by the coefficient of variation of the group variance, c. According to Box (1954), although c can be as large as $(K-1)^{\frac{1}{2}}$, values greater than one are extremely rare in practice. Therefore, four degrees

of variance heterogeneity were investigated, corresponding to c values of .2, .4, .6, and 1.0. Table 1 gives the values of the group variances for each value of c.

Table 1

Group Variances for Given Coefficients of Variation

c	Group 1	Group 2	Group 3	Group 4
.2	.7317	.9106	1.0894	1.2683
.4	.4633	.8211	1.1789	1.5367
.6	.1950	.7317	1.2683	1.8050
1.0	.1515	.4343	.7172	2.6971

Power of the tests of mean equality was investigated by simulating conditions of unequal group means. This was achieved by the addition of an appropriate constant, i.e., treatment effect, to each of the observations within each of the K levels of the treatment variable. The size of the treatment effects were chosen so as to give an intermediate "effect size" as defined by Cohen (1969). Effect size, f, is given by:

$$f = \frac{\sigma_{\mu_j}}{\sigma_e} , \qquad (40)$$

where $\sigma_{\mu_j} = \sqrt{\Sigma p_j \ (\mu_j - \mu)^2}$,

$$p_j = n_j/N \quad ,$$

$$N = \sum_j n_j \quad \text{and}$$

$$\mu = \sum_j n_j \mu_j / N \quad .$$

f = .25 is considered by Cohen (1969) to be a value representative of intermediate sized effects found in behavioural research and is the value which was approximated here.

As was pointed out previously, when variances and means differ, the way in which these two parameters are paired affects the power of the test of mean differences. Also, the different tests are affected in different ways by a given combination of means and variances. Thus, in order to thoroughly investigate this phenomenon, two strategies were adopted: firstly, means and variances were both positively and negatively correlated across groups, i.e., as the means increased from groups one to four, the variances also increased (positively correlated) or as the means increased from groups one to four, the variances decreased (negatively correlated); and, secondly, two patterns of mean differences were investigated. Group means were either dichotomized at each end of the range of means or were spaced equidistantly from each other over the range, i.e., the pattern of means was either $\mu_1 = \mu_2 < \mu_3 = \mu_4$ or $\mu_1 < \mu_2 < \mu_3 < \mu_4$, where each successive mean increases by a constant amount. These two patterns of mean differences do not give rise to the same effect size, if the range of the means is constant, since dichotomizing the means leads to greater mean variability (σ_{μ_j}) than spreading them

equidistantly over the range. In order to obtain an equal effect size for both patterns, the range of the equidistant means must be greater than that of the dichotomized means. Table 2 gives the values of the means which were used: these values were determined for f = .25, $\sigma_e = 1$, $\mu = 0$ and equal sample sizes.

Table 2

Group Means[a]

Pattern	Group 1	Group 2	Group 3	Group 4
Equidistant	-.3354	-.1118	+.1118	+.3354
Dichotomized	-.2500	-.2500	+.2500	+.2500

[a]Entries in the table are based on a Cohen's (1969) f = .25, $\sigma_e = 1$ and $\mu = 0$.

Since all populations start out with a mean of zero, approximately 50 per cent overall, of the individual generated observations had a negative value. (This was true even when treatment effects were present as the mean of the means, $\mu = \Sigma_j \mu_j / K$, was still held at zero under these circumstances.) This posed a problem when transforming the data prior to an ANOVA F-test, as the built-in computer functions for determining logarithms and square roots cannot accept negative numbers. In order to eliminate negative values, a sufficiently large positive constant (10) was added to all generated observations prior to all transformations.

Group sizes were chosen to give an a priori ANOVA F-test power of approximately .70. The values for the group means in Table 2 give an effect size of .25, when the group sizes are equal. According to Cohen's (1969, p.377) tables, a group size of 36 is required to give a power of .70 for f = .25. When group sizes are unequal, this disparity was quantified by the ratio of the maximum to minimum group size (see Spjøtvoll and Stoline, 1973). Thus, u, the degree of sample size imbalance, is given by:

$$u = \max(n_1,\ldots n_K) \;/\; \min(n_1,\ldots n_K) \quad .$$

Three levels of u were used: small, u = 1.4 or 1.5; medium, u = 2.0; and large, u = 3.0. The same total N was used when the n_j's were unequal and the group sizes were spread approximately equidistantly over the range between $\max(n_j)$ and $\min(n_j)$. However, as may be seen from inspection of equation (41), introduction of unequal sample sizes alters the value of σ_{μ_j} for given values of the group means. If the groups whose means are extreme, i.e., have large $(\mu_j - \mu)^2$, also have large n_j's relative to the others, f, and therefore power, is larger than with equal n_j's; conversely, if extreme groups have small n_j's, f and power is smaller (see Cohen, 1969, p.353. Therefore, different values of u for the same group mean values and total N lead to different values for power. Thus the empirically determined power values were expected to deviate from the value of .70 when group sizes are unequal. The upper part of Table 3 gives the group sizes for different u values when power is .70 for the equal n_j condition: f values for

each condition are also shown. It is not possible to accurately determine power for these f values either from Cohen's (1969) tables or from the standard nomographs for ϕ ($\phi = f\sqrt{n}$): however for a mean n_j of 36 the power ranges from a value of .50 at f = .20 to .70 at f = .25.

As can be seen, choosing group sizes on the basis of power considerations results in large values per group, if Cohen's (1969) intermediate effect size is used. Unfortunately, however, E may not always be able to obtain such large groups and the power of his tests of mean differences will, therefore, suffer. Under these circumstances the performance of a variance test in making decisions between alternative tests of mean equality may become crucial. Therefore, smaller group sizes comparable to those used in other studies (e.g., Kohr and Games, 1974, and Martin and Games, 1976) were also investigated. When the total N was 48, the lower part of Table 3 gives the group sizes for the equal n_j condition and the various degrees of group size imbalance. It is worthy of note that, at a mean n_j of 12, the power associated with f ranges from only .26 at f = .25 to .17 at f = .20.

Table 3

Group Sizes and f Values

u	f(em[a])	f(dm[b])	Group 1	Group 2	Group 3	Group 4
		N chosen to give a power of ≃ .70[c]				
1.0	.250	.250	36	36	36	36
1.5	.247	.248	29	34	38	43
2.0	.242	.244	24	32	40	48
3.0	.232	.236	18	30	42	54

Table 3 (cont.)

u	f(em[a])	f(dm[b])	Group 1	Group 2	Group 3	Group 4
	Mean n_j chosen to equal Martin and Games (1976) mean n_j					
1.0	.250	.250	12	12	12	12
1.4	.248	.248	10	11	13	14
2.0	.243	.245	8	11	13	16
3.0	.232	.236	6	10	14	18

[a]em = equidistant means

[b]dm = dichotomized means

[c]f = .25

α = .05

When variances and group sizes differ across groups, these two quantities were also both positively and negatively paired with each other in order to simulate conditions giving rise, respectively, to conservative and liberal F-tests. Values of the Box (1954) bias coefficient, b, for the combined unequal group size and variance conditions are given in Table 4. The value of b varies from .6714 to 1.6522, an extensive range of bias.

Figure 1 summarizes all of the 193 combinations of mean, variance and group size variability conditions that were simulated. All these conditions were generated for two levels of sample size: one chosen to give an a priori power of .70 and the other chosen to match the sample size used by Martin and Games (1976). (Values of the Box (1954) bias coefficient are only slightly more extreme under the latter conditions and are, therefore, not tabled.) This whole procedure will

Table 4

Values of the Box Bias Coefficient for Given Group Size and Variance Inhomogeneity

Unweighted Coefficient of Variation	Group Size Imbalance	Pairing of Group Sizes and Variances[a]	Box Bias Coefficient
c	u		b
.2	1.5	POS	.9632
.4	"	"	.9284
.6	"	"	.8954
1.0	"	"	.8501
.2	"	NEG	1.0389
.4	"	"	1.0803
.6	"	"	1.1242
1.0	"	"	1.1938
.2	2.0	POS	.9397
.4	"	"	.8803
.6	"	"	.8282
1.0	"	"	.7612
.2	"	NEG	1.0692
.4	"	"	1.1461
.6	"	"	1.2320
1.0	"	"	1.3735
.2	3.0	POS	.9081
.4	"	"	.8282
.6	"	"	.7580
1.0	"	"	.6714
.2	"	NEG	1.1066
.4	"	"	1.2320
.6	"	"	1.3814
1.0	"	"	1.6522

[a]POS - positive pairing of group sizes and variances
NEG - negative pairing of group sizes and variances

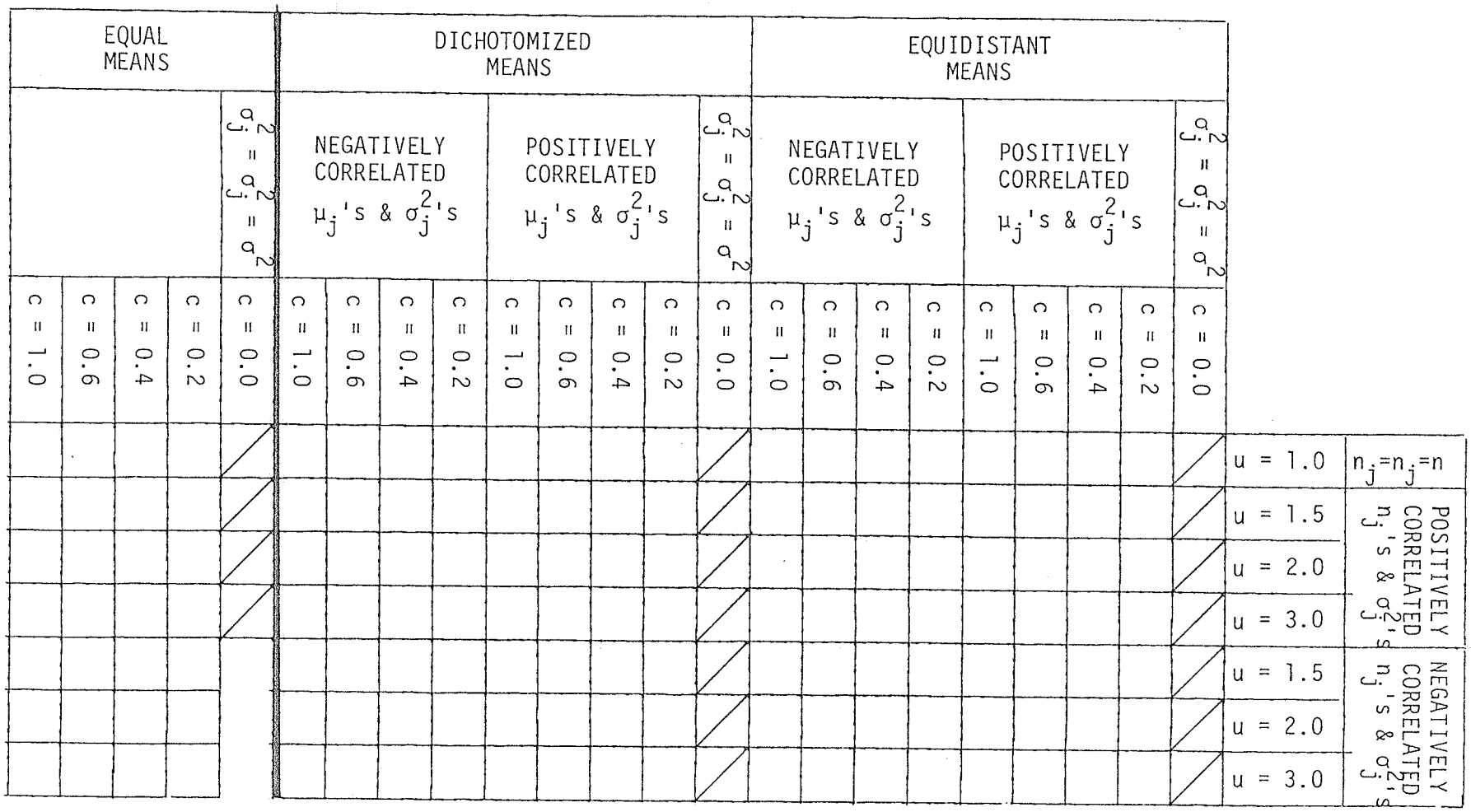


Figure 1. Diagrammatic representation of the conditions which were investigated for each of the four combinations of total N and population shape.

be repeated for each of the populations (normal and χ_2^2) giving rise to a total of 2 x 2 x 193 = 772 conditions.

A total of 2000 sets of data were generated for each of the 772 conditions. All the individual tests and test combinations under the headings A, B and C above were calculated on each of the 2000 data sets; then, for each of the tests or combination of tests, the proportion of the 2000 that were significant were recorded.

Referring again to Figure 1, it can be seen that all cells below the heavy horizontal line represent conditions in which the group means were equal; therefore, the proportions of significant means test statistics recorded in these cells represent the probability of a Type I error for the means tests. On the other hand, all cells above the horizontal line represent conditions of group mean inequality and, therefore, proportions of significant means test statistics recorded here represent the power of the means tests.

The cells of Figure 1 which contain a diagonal line represent conditions of homoscedasticity and, therefore, proportions of significant variance test statistics recorded in these cells give the probability of a Type I error for the variance tests. Cells in the remainder of Figure 1 all represent heteroscedastic conditions and, therefore, the proportion of significant variance test statistics recorded for them give the power of the variance tests.

During the "debugging" of the computer program, which calculated all the statistics of the tests investigated, a negative value was calculated for Box and Andersen's (1955) M'. Checking of all the

computational steps revealed no errors but a value below -2.0 was found for c_2, the correction factor in the denominator of M' (see p.34), thus accounting for the obtained negative value.

Since the formula for M' seemed to be predicated upon the fact that c_2 should never approach -2.0, let alone become less than this value, a sampling study of c_2 was conducted. In all cases four groups of 18 normally distributed variates were generated and, if desired, variance heterogeneity was introduced as described above: c_2 was then computed. This procedure was repeated 1000 times for each level of variance heterogeneity investigated thus producing a sampling distribution of 1000 cases for c_2. Table 5 gives the results obtained.

Table 5

Highest and Lowest c_2 values and percentages of c_2 values less than -2.0

Population Variances				Lowest c_2	Highest c_2	$\%c_2 < -2.0$
Group 1	Group 2	Group 3	Group 4			
1.0000	1.0000	1.0000	1.0000	-1.4219	3.2468	0
.4633	.8211	1.1789	1.5367	-1.9367	3.7488	0
.1515	.4343	.7172	2.6971	-4.3821	12.3247	7.4

These results indicate a fairly high probability of obtaining a negative value of M' (or $c_2 < -2.0$) if the highest degree of variance heterogeneity (i.e., c = 1.0) is present.

Since Bartlett's (1937) test performs very well on normal and platykurtic populations, a possible solution to the problem of c_2 values less than -2.0 seemed to be using M if c_2 is less than zero and M' if c_2 is greater than zero. This combined procedure has been designated "combined M", and was added to the list of variance tests investigated. Thus two more sequential procedures were also generated by using "combined M' to choose between either the ANOVA F and Welch test or between the ANOVA F and F* tests. (These procedures are designated as FW/CM and FF*/CM, respectively.)

Results and Discussion

Type I Errors of the Means Tests

Type I error rates of the means tests are presented in Table 6 for the small sample size. When the population sampled had a normal distribution, the ANOVA F-test showed its familiar characteristics of (a) becoming liberal when group sizes and variances were unequal and larger group sizes were paired with smaller variances, and (b) becoming conservative when the relationship between group sizes and variances was reversed. In contrast, the Welch test showed its usual excellent control of Type I error rates regardless of the degree of assumption violation. Overall the Brown and Forsythe F* test was slightly liberal, never exceeding the nominal level of significance by more than 2.85 percentage points.

When sampling from the non-normally distributed (χ_2^2) population the Type I error rates for the ANOVA F and F* tests were not markedly different than those obtained for the normal population. In contrast, the Welch test showed an increase in Type I error rates for most conditions investigated. Not only was the nominal alpha level exceeded by as much as 8.4 percentage points for the more extreme cases of positive bias, but also, the Welch test performed worse than the ANOVA F-test, when a small degree of positive bias was present (i.e., a Box bias coefficient between 1.0 and 1.3). Nominal alpha levels were also exceeded by this test in the presence of negative bias and when only variances or group sizes were heterogeneous.

Table 6.

Empirical Type I error rates (%) for the means tests (small N)

Group sizes condition	Coefficient of variation	Normal distribution			Chi-square distribution		
		F	W	F*	F	W	F*
u = 1.0	0.0	5.25	5.05	5.05	3.80	4.45	3.25
	0.2	5.15	5.30	4.90	4.50	5.60	3.95
	0.4	6.15	4.75	5.70	5.20	6.30	4.30
	0.6	5.85	5.55	5.40	6.25	9.20	5.30
	1.0	8.50	5.25	6.85	9.55	8.85	8.30
Positively correlated group sizes and variances							
u = 1.5	0.0	4.10	4.30	3.90	4.35	5.90	4.00
	0.2	5.20	5.15	5.75	4.65	5.80	4.60
	0.4	4.30	5.10	5.15	4.10	5.95	4.25
	0.6	5.45	5.65	6.35	5.80	9.85	6.15
	1.0	5.85	4.55	6.90	7.80	9.00	8.90
u = 2.0	0.0	4.55	5.05	4.90	4.30	5.65	3.75
	0.2	4.05	4.85	4.70	4.10	5.50	4.20
	0.4	3.15	4.80	4.60	4.45	5.70	4.85
	0.6	3.40	4.30	5.60	3.60	7.50	4.95
	1.0	3.95	4.80	6.05	5.70	7.65	7.85
u = 3.0	0.0	5.00	5.45	4.65	4.55	7.65	3.95
	0.2	4.20	5.60	5.30	3.95	6.80	4.45
	0.4	2.80	4.70	5.50	3.35	5.30	4.55
	0.6	2.85	4.90	5.95	3.65	7.75	5.70
	1.0	3.25	5.75	7.55	4.25	6.70	7.80
Negatively correlated group sizes and variances							
u = 1.5	0.2	5.15	4.80	4.55	4.75	5.25	3.95
	0.4	6.50	5.35	5.50	6.20	7.20	4.25
	0.6	8.85	5.45	6.05	7.05	10.00	4.35
	1.0	11.35	5.35	7.45	12.80	11.35	8.80
u = 2.0	0.2	6.55	5.40	5.30	4.70	6.70	3.80
	0.4	8.75	5.50	5.60	8.45	9.90	5.20
	0.6	10.25	4.50	5.90	11.00	12.35	7.30
	1.0	15.85	5.15	7.85	16.00	10.85	8.70
u = 3.0	0.2	5.75	4.25	3.75	5.50	8.15	4.40
	0.4	9.60	5.80	4.85	10.05	12.30	6.35
	0.6	13.55	5.25	5.65	13.65	13.40	5.95
	1.0	20.70	6.40	7.25	22.40	12.70	10.30

A comparison of the differences in performance of the Welch and F* tests indicates that the former is more robust for normal distributions and the latter for non-normal distributions. Since the difference in Type I error rates was generally greater in the non-normal population, this would indicate a preference for the F* test when no prior information is available regarding the shape of the population distribution: obviously, if the distribution form is known, this should dictate the choice of test.

Table 7 shows the Type I error rates of the means tests for the large sample size (N=144). Essentially the same pattern of results was obtained for each test at both sample sizes but there were differences between tests in the response to increasing total sample size. The major difference was the improved performance of the Welch test in the chi-square population: since this test was the one most affected by non-normality it is not surprising that it showed improvement on increasing sample size. In contrast the F* test became slightly more liberal, overall, at the larger N (the maximum deviation from nominal alpha now being 3.35%). Because of the opposite effect of sample size on these two tests, the case for preferring the Brown and Forsythe F* test, although still extant, was not so convincing.

Power of the Means Tests

Power at small sample size. Power values obtained for the means tests using the small sample size (N=48) are presented in Table 8. Because of the extent of the data, results for the equal and most unequal group size conditions, only, are presented. Results obtained for the

Table 7.

Empirical Type I error rates (%) for the means tests (large N)

Group size condition	Coefficient of variation	Normal distribution			Chi-square distribution		
		F	W	F*	F	W	F*
u = 1.0	0.0	4.80	5.05	4.80	5.00	5.85	4.75
	0.2	5.90	5.95	5.90	4.90	6.25	4.75
	0.4	5.80	5.45	5.65	6.25	6.00	6.15
	0.6	6.35	5.50	6.30	5.80	5.90	5.55
	1.0	7.75	4.80	7.35	7.95	6.45	7.75
Positively correlated group sizes and variances.							
u = 1.5	0.0	4.95	5.15	4.80	4.00	5.45	4.15
	0.2	4.55	5.00	5.65	4.55	5.65	4.70
	0.4	5.60	5.60	6.95	5.15	6.05	5.85
	0.6	5.50	5.85	7.30	5.15	6.75	6.70
	1.0	5.75	5.75	8.35	6.10	6.80	7.75
u = 2.0	0.0	4.50	4.85	4.55	4.80	5.95	4.55
	0.2	3.55	3.50	4.25	4.00	6.20	5.75
	0.4	4.30	5.90	6.40	3.45	5.40	5.40
	0.6	2.65	3.85	5.05	3.35	6.50	5.75
	1.0	3.10	4.50	7.05	4.80	6.25	7.30
u = 3.0	0.0	4.75	5.05	4.90	5.30	7.05	4.70
	0.2	3.90	5.15	5.40	4.00	6.20	5.75
	0.4	3.10	5.20	5.85	2.50	5.10	5.15
	0.6	2.90	4.90	6.15	2.65	5.30	5.85
	1.0	3.10	5.90	7.75	3.40	6.25	8.15
Negatively correlated group sizes and variances.							
u = 1.5	0.2	6.00	5.80	5.55	5.60	5.80	4.45
	0.4	7.25	5.40	5.80	6.65	6.30	5.10
	0.6	8.45	5.35	6.45	7.80	6.35	5.65
	1.0	10.35	4.90	6.70	12.30	7.75	8.90
u = 2.0	0.2	7.00	5.30	5.85	6.75	6.90	5.65
	0.4	6.80	4.30	4.25	7.85	7.30	5.45
	0.6	10.50	5.45	5.75	10.40	8.30	5.95
	1.0	14.95	4.95	6.95	16.25	8.75	9.00
u = 3.0	0.2	8.50	5.80	6.00	5.70	6.10	4.10
	0.4	10.05	5.00	5.60	9.95	7.20	5.70
	0.6	13.20	4.90	5.85	14.90	10.65	7.20
	1.0	20.30	5.30	7.70	20.95	10.30	9.65

Table 8.

Empirical Power Values for the Means Tests (Small N)

Means Condition	Correlation of Means & Variances	Coefficient of Variation	Normal Distribution F	Normal Distribution W	Normal Distribution F*	Chi-Square Distribution F	Chi-Square Distribution W	Chi-Square Distribution F*
Equal group sizes								
Equi-distant	positive	0.0	28.05	25.50	27.45	29.35	30.65	27.50
		0.2	26.45	25.85	25.95	23.95	26.70	22.45
		0.4	24.95	27.15	23.80	23.80	28.15	21.60
		0.6	22.55	28.50	21.30	20.35	23.90	17.40
		1.0	26.00	36.05	23.15	19.95	28.40	16.35
	negative	0.0	28.05	25.50	27.45	29.35	30.65	27.50
		0.2	25.30	24.55	24.65	30.25	32.05	29.10
		0.4	24.95	26.05	24.15	30.50	38.50	29.35
		0.6	24.75	31.00	23.70	33.35	50.15	31.55
		1.0	24.60	34.30	21.80	34.20	53.10	32.10
Dichot-omized	positive	0.0	26.90	25.65	26.55	27.40	30.90	25.65
		0.2	25.35	24.80	24.75	28.30	29.30	26.80
		0.4	24.75	23.65	24.00	24.45	25.50	22.75
		0.6	26.50	26.05	24.20	23.60	20.75	20.50
		1.0	25.70	33.95	23.20	21.75	32.35	17.65
	negative	0.0	26.90	25.65	26.55	27.40	30.90	25.65
		0.2	21.75	22.00	21.10	30.15	35.60	28.85
		0.4	25.90	24.90	25.25	32.05	37.45	30.75
		0.6	24.80	25.60	22.90	34.25	43.60	32.45
		1.0	25.20	32.35	21.90	34.20	49.85	31.25
Positively correlated group sizes and variances (u = 3.0)								
Equi-distant	positive	0.0	22.45	20.85	21.95	23.35	36.75	26.70
		0.2	19.45	21.95	22.55	17.30	33.80	24.90
		0.4	14.20	23.50	21.85	12.45	33.50	22.60
		0.6	13.75	32.05	24.00	10.15	34.65	21.30
		1.0	11.35	32.55	24.10	6.35	34.75	18.40
	negative	0.0	19.80	19.45	19.30	24.20	17.70	18.50
		0.2	18.20	21.50	21.70	23.35	22.25	21.80
		0.4	16.70	25.95	24.05	20.25	28.70	23.75
		0.6	12.35	30.45	21.60	18.15	43.90	26.30
		1.0	10.90	32.70	22.80	21.30	46.40	30.60
Dichot-omized	positive	0.0	23.15	20.40	22.10	25.30	36.15	26.50
		0.2	16.60	20.20	20.60	18.90	34.95	25.90
		0.4	15.00	23.95	22.35	13.00	34.25	23.70
		0.6	13.00	27.45	23.25	9.75	31.65	22.95
		1.0	10.80	34.10	23.65	7.30	41.85	20.75
	negative	0.0	22.05	21.05	21.30	26.60	20.45	20.15
		0.2	18.30	21.65	21.80	24.40	22.85	23.50
		0.4	15.80	25.10	23.75	22.30	29.30	25.75
		0.6	12.95	26.85	24.55	19.50	38.65	27.65
		1.0	10.10	32.55	22.70	19.60	45.30	30.80
Negatively correlated group sizes and variances (u = 3.0)								
Equi-distant	positive	0.0	19.80	19.45	19.30	24.20	17.70	18.50
		0.2	25.70	21.00	19.05	27.85	15.45	15.25
		0.4	31.65	21.35	19.55	28.95	12.80	12.60
		0.6	36.00	24.00	18.60	34.00	12.25	11.60
		1.0	41.65	28.20	16.55	39.75	19.10	11.30
	negative	0.0	22.45	20.85	21.95	23.35	36.75	26.70
		0.2	26.35	20.85	20.05	31.15	38.50	29.05
		0.4	32.70	22.20	19.70	37.45	42.90	29.95
		0.6	36.15	23.20	19.25	43.90	47.85	29.65
		1.0	45.20	31.75	18.60	51.45	52.75	31.35
Dichot-omized	positive	0.0	22.05	21.05	21.30	26.60	20.45	20.15
		0.2	26.25	19.85	20.30	26.95	16.80	17.25
		0.4	30.20	18.40	19.10	28.95	13.30	13.20
		0.6	39.05	19.65	19.75	34.20	12.25	12.95
		1.0	45.95	25.90	18.55	46.30	20.90	15.60
	negative	0.0	23.15	20.40	22.10	25.30	36.15	26.50
		0.2	28.75	21.05	21.15	30.60	38.40	28.40
		0.4	31.25	19.10	18.20	37.95	38.55	29.10
		0.6	36.50	19.45	18.70	45.75	43.20	29.25
		1.0	45.45	26.45	17.85	52.95	48.00	30.50

small and intermediate levels of group size inequality show the same pattern as those presented in Table 8 (for the greatest degree of group size inequality) but effects are not as marked.

Confining attention firstly to the normal distribution results, it will be seen that neither the pattern of mean differences (i.e., equidistant vs. dichotomized) nor the pairing of group means and variances had any appreciable effects upon the power of the tests.

As was anticipated on a priori grounds, the power of the ANOVA F-test was about 5% less in the extremely unequal group size condition (u=3.0) than when n_j's were equal (for equal group variances). Referring back to page 63, it will be seen that the power values for F fall in the expected range. Both W and F* were similarly affected.

Violation of the homoscedasticity assumption in the presence of equal group sizes had a relatively small effect on F and F* and a somewhat larger effect on W. The ANOVA F-test and the F* test were affected almost identically, both showing a small decrease in power with increasing variance heterogeneity, whereas the Welch test showed a slightly larger increase ($\sim$ 10%). Thus the tests have about equal power when variances are equal but as heteroscedasticity increases the power superiority of the Welch test becomes more pronounced.

With the introduction of unequal group sizes, heteroscedasticity produced more marked effects on the power of F, about the same magnitude of effects on the power of W and negligible effects on F*. When larger group sizes occurred in groups having larger variances, the power of F dropped from $\sim$ 23% to $\sim$ 10% as c (the coefficient of variation of the

variances) increased, and, when the pairing of group sizes and variances was reversed, power of F rose from ∿ 20% to ∿ 42% between zero and maximum variance heterogeneity. For W, the pairing of group sizes and variances made only a small difference to the effect of variance heterogeneity, the effect being greater for the positive pairing. For both pairings there was an increase in power, as when the group sizes were equal. Overall, then, the rankings of the tests for power were W > F* >F for positively paired group sizes and variances and F > W > F* for the negative pairing.

Responses of W and F* to increasing variance heterogeneity were not substantially influenced by the degree of group size inequality, but, as may be anticipated from the calculated bias coefficients, increasing values of u led to an increasing responsiveness to heteroscedasticity for F. These relationships are illustrated in Figure 2.

Turning now to the results obtained when sampling from the chi-square distribution and confining attention to the equal group size data, two distinct differences emerge between these results and the normal distribution data. Firstly, the response to variance heterogeneity is different for the two pairings of means and variances and, secondly, the chi-square values are generally larger. As for the normal distribution, the pattern of results differs negligibly for the two patterns of dispersion of the group means (see Figure 3).

When group sizes are equal and larger group means are associated with larger variances, both F and F* show a power drop of ∿ 10% over the range of increasing c values, while W shows a smaller drop. Reversal

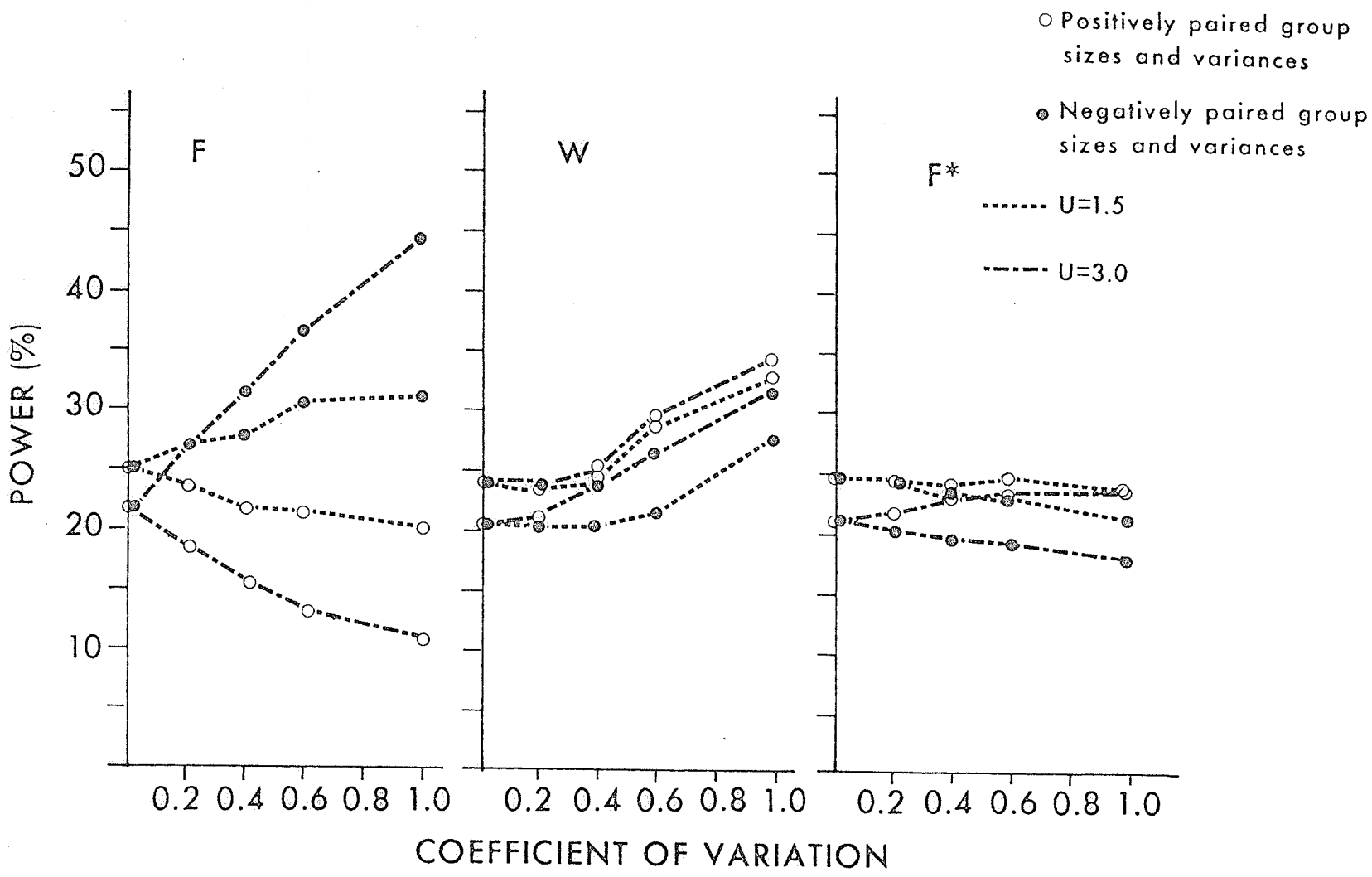


Figure 2. Power of the means tests as a function of variance heterogeneity, group size variance pairing and group size inequality. (Normal distribution - small N - data points averaged over pattern of mean differences and mean variance pairing.)

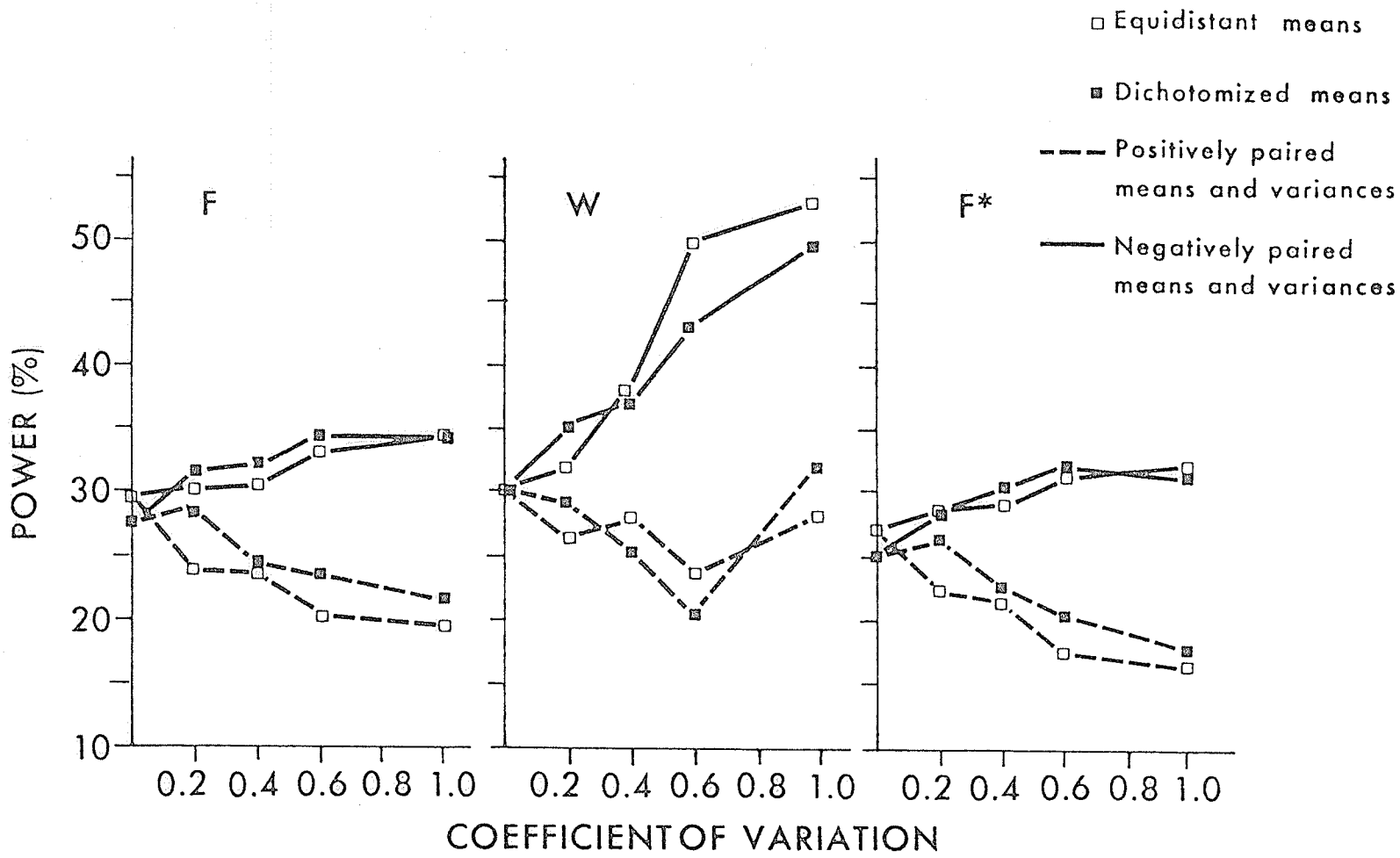


Figure 3. Power of the means tests as a function of variance heterogeneity, pattern of mean differences and mean variance pairing. (Chi-square distribution - small N - equal group sizes.)

of the pairing of means and variances leads to a slight power increase for F and F* (∿5%) as heteroscedasticity increases and a dramatic increase from ∿ 31% to ∿ 53% for W. Overall for the equal group size condition, there is little to choose between the power of F and F* and that of W is superior to both, especially when means and variances are negatively paired. Figure 3 illustrates the effects of mean-variance pairing, when group sizes are equal, for all three tests.

When group sizes are unequal, the pairings of group sizes with variances and of means with variances automatically results in specific pairings of means and group sizes. Thus, when means and group sizes are either both positively or both negatively paired with variances, they are also positively paired with each other, and, when the pairing of means and variances is in the opposite direction to the pairing of group sizes and variances, the means and group sizes are negatively paired with each other. The result of these relationships between group sizes, variances and means is that any interaction of the effects of mean-variance and group size-variance pairing will necessarily be confounded with the effect, if any, of mean-group-size pairing. Only when variances are homogeneous can the effect of mean-group-size pairing be evaluated. Thus results for this effect may be found in Table 8 under either positively or negatively correlated group sizes and variances when c=0. While the relationship of means and group sizes appeared to be irrelevant in determining power when sampling from the normal distribution, it became a sizeable factor for the chi-square distribution.

For the ANOVA F-test extremely unequal group sizes produced

about the same power loss (from the equal group sizes condition) as occurred with the normal distribution, regardless of the pairing of means and group sizes. In contrast, the Welch test showed a power increase of $\sim$ 6%, when means and group sizes were positively paired, and, a decrease of 10%-13% for the negative pairing of means and group sizes. F* had similar power to the equal group size condition, for positive pairing of means and group sizes, but, when the pairing was reversed, there was a $\sim$ 9% drop.

In the presence of extremely unequal group sizes and heterogeneous variances there were three factors influencing the power of the means tests, in the data presented in Table 7, namely, degree of variance heterogeneity, pairing of means and variances and pairing of means and group sizes. Inspection of Figure 4 shows that while there were similarities in the interactions of these factors on W and F*, the pattern of interactions was different from F. The ANOVA F-test demonstrated the expected interaction between variance heterogeneity and pairing of group sizes and variances, and when means and variances were positively paired this interaction was practically identical in form to that which occurred with the normal distribution for both pairings of means and variances. However, when means and variances were negatively paired, although the interaction between variance heterogeneity and group size-variance pairing was almost the same size, increasing variance heterogeneity caused practically no drop in power for negatively paired group sizes and variances and a much larger increase for positively paired group sizes and variances: thus there was also a clear interaction between degree of

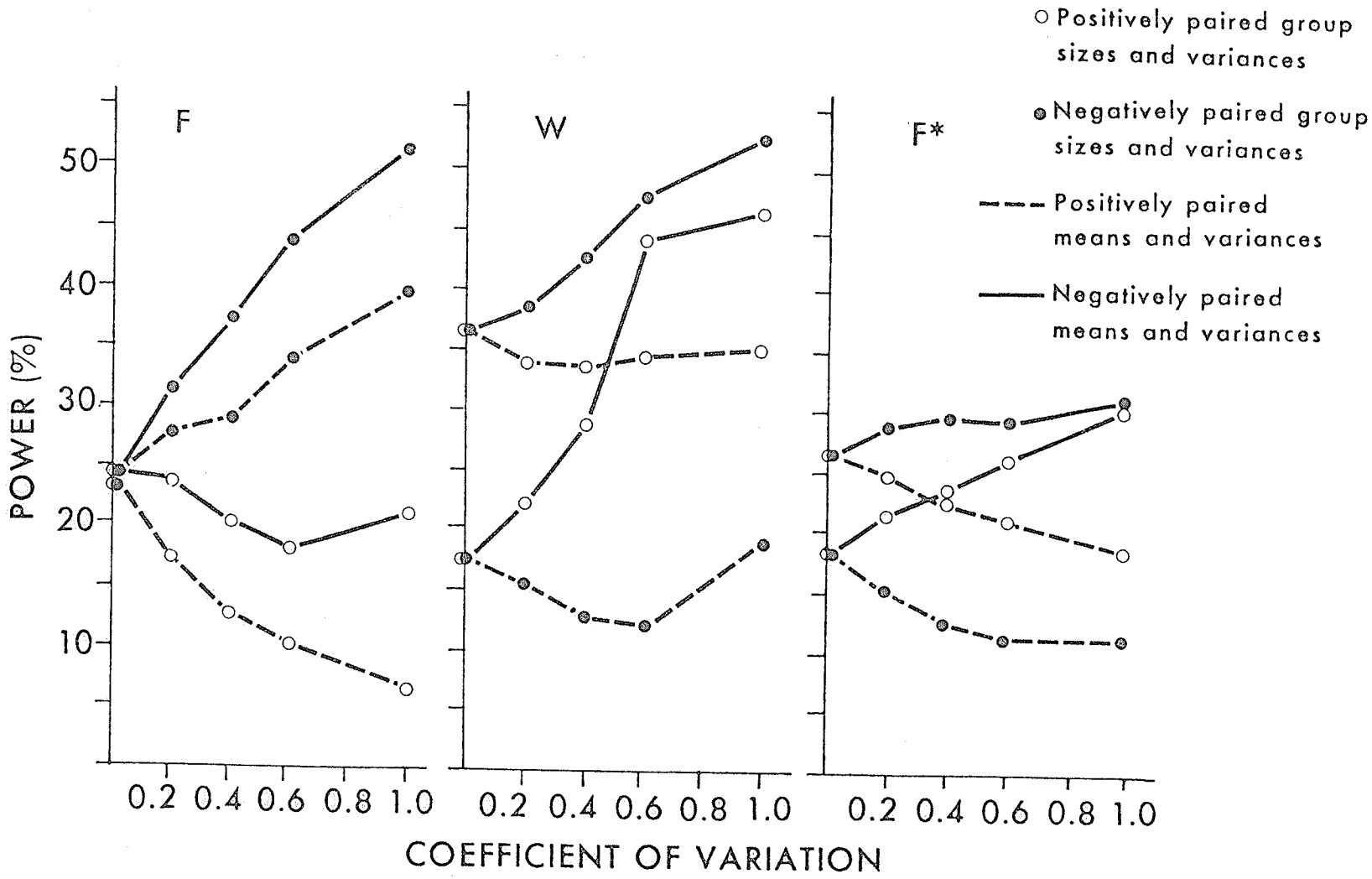


Figure 4. Power of the means tests as a function of variance heterogeneity, mean-variance pairing and group size variance pairing. (Chi-square distribution - small N - equidistant means - u = 3.0.)

variance heterogeneity and mean-variance pairing. Since the latter interaction was identical over both pairings of group sizes and variance, there was obviously no interaction between mean-variance pairing and group size-variance pairing and no triple interaction between the three factors affecting power. In summary, the most important difference between the results for F in the two population distributions was the presence of an interaction between degree of variance heterogeneity and mean-variance pairing in the chi-square distribution and its absence in the normal distribution. Apparently, for the chi-square population distribution, negative pairing of means and variances causes increasing power of the F-test as variance heterogeneity increases, while the reverse is true for positive mean-variance pairing, and, this effect of mean-variance pairing is additive with the effects on power of group size-variance pairing.

As stated above, the three factors of variance heterogeneity, mean-variance pairing and group size-variance pairing combined to produce similar effects on the power of the Welch and F* tests. A triple interaction existed between these three factors in their effects on the two tests with the Welch test being most affected by the three manipulations. When means and variances were positively paired, there was little effect of variance heterogeneity on the power of W for either pairing of group sizes and variances. If anything, a bowed relationship existed between variance heterogeneity and power, i.e., power fell then rose again with increasing c, the maximum difference between any two c values being $\sim$ 7%: this relationship was the same for both

pairings of group sizes and variances, the same difference in power levels of ∿ 20% existing for all values of c. If power is averaged over pattern of mean differences and variance heterogeneity, for this level of sample size inequality, it is ∿ 35% for negative group size-variance pairing and ∿ 15% for positive group size-variance pairing. Presumably this difference is due to the pairing of means and group sizes only, since it exists when homoscedasticity prevails.

For F* much the same effects were observed, except that the pattern of the relationship between power and variance heterogeneity was a steady decline as c increased, but again, the maximum difference between any two values of c was ∿ 7%. However, the averaged power, when variances and group sizes were positively paired, was ∿ 23%, and, for the opposite relationship of variances and group sizes it was ∿ 14% thus giving a difference of only ∿ 9%, which was considerably less than for the Welch test.

When means and variances were negatively paired, both W and F* showed increases in power with increasing variance heterogeneity, this relationship being steeper for the positively paired group sizes and variances for both tests. Averaging over patterns of mean differences, the power increases from c=0 to c=1.0 for W and F* were ∿ 27% and ∿ 11%, respectively, for positively paired group sizes and variances and ∿ 14% and ∿ 4%, respectively, for the negative group size-variance pairing.

The interaction of variance heterogeneity (C), mean-variance pairing (V), and group size-variance pairing (S) on the power of W and F* can best be visualized, by noting in Figure 4, the completely different pattern of VC interaction for the two different pairings of group

sizes and variances.

As for the normal population, the degree of group size inequality greatly influenced the response to variance heterogeneity of F, but, unlike the results for the normal distribution, there was also a substantial effect on W and a small effect on F*. Figure 5 illustrates the results obtained when the degree of group size inequality was smallest and a comparison of Figures 4 and 5 demonstrates the effects of group size inequality on all three tests. From the graph for the Welch test it can be seen that the previously discussed three-way interaction (SxVxC) was somewhat less when group size inequality (u) was at its smallest value but the most dramatic effect of reducing u was on the two-way interaction between mean-variance pairing and group size-variance pairing (VxS): this latter effect is illustrated more clearly in Figure 6. Again, it must be remembered that any VxS interaction is confounded with the effects of group size-mean pairing. From Figures 4 and 5, it appears that the effect of mean and group size pairing, per se (i.e., when variances are equal), is about three times larger for the greater degree of group size inequality and, from Figure 6, this is also the ratio of the VxS effects at the two levels of u. Similar, though much smaller effects are apparent for F*.

Power at large sample size. Results for the large sample size (N=144) are presented in Table 9. For the normal population distribution the effects of the various manipulations were in the same direction only larger. Whereas, for the small sample size, the combined effects of variance heterogeneity, group size inequality and group size-variance

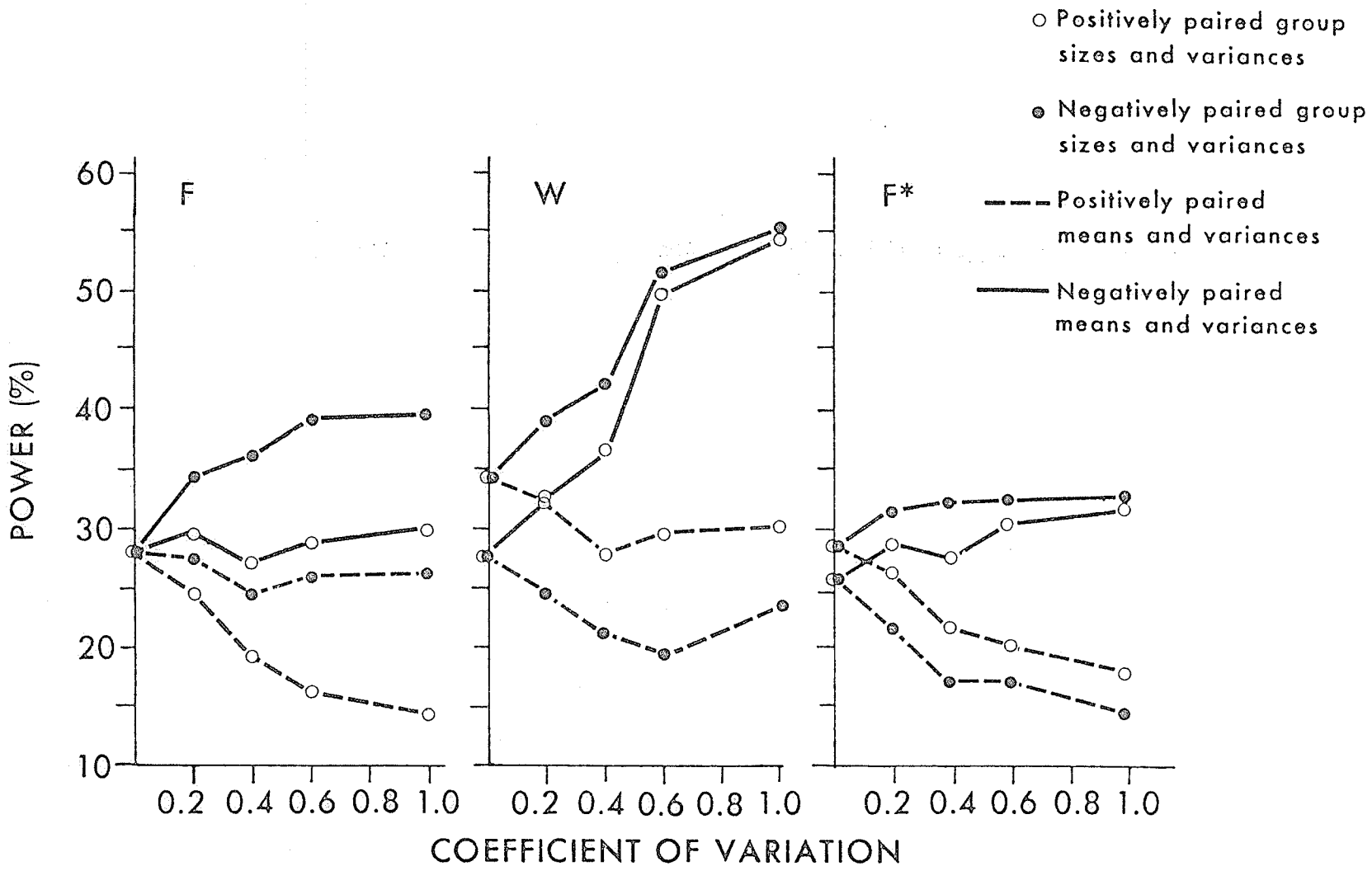


Figure 5. Power of the means tests as a function of variance heterogeneity, mean-variance pairing and group size variance pairing. (Chi-square distribution - small N - equidistant means - u = 1.5.)

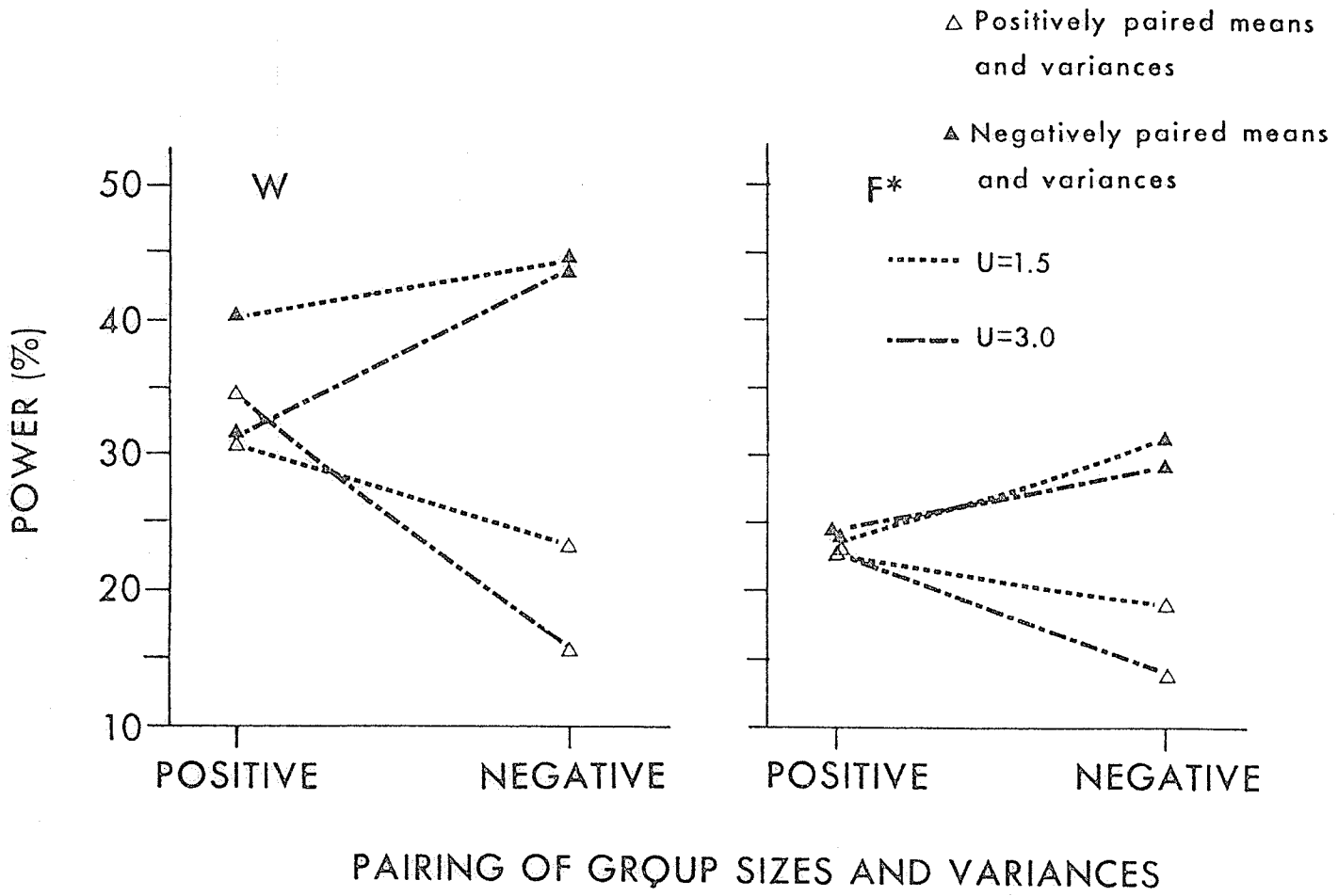


Figure 6. Power of the Welch (W) and F* tests as a function of group size-variance pairing, mean-variance pairing and group size inequality. (Chi-square distribution - small N - equidistant means - data points averaged over degrees of variance heterogeneity.)

Table 9.

Empirical Power Values for the Means Tests (Large N)

Means Condition	Correlation of Means & Variances	Coefficient of Variation	Normal Distribution			Chi-Square Distribution		
			I	W	I*	I	W	F*
Equal group sizes								
Equi-distant	positive	0.0	70.55	69.20	70.50	71.10	71.05	71.00
		0.2	68.35	68.55	68.10	71.40	71.00	71.20
		0.4	68.75	74.80	68.45	72.50	75.15	72.20
		0.6	69.10	82.25	68.30	73.95	84.50	73.30
		1.0	64.20	87.30	62.75	70.00	88.20	68.30
	negative	0.0	70.55	69.20	70.50	71.10	71.05	71.00
		0.2	70.95	70.50	70.90	68.95	72.35	68.85
		0.4	70.00	74.20	69.70	68.50	77.15	68.15
		0.6	67.25	81.70	67.05	68.25	85.40	67.50
		1.0	65.40	87.15	64.10	65.80	90.80	64.45
Dichot-omized	positive	0.0	68.30	67.45	68.30	70.05	70.15	71.00
		0.2	69.65	68.80	69.60	71.55	72.35	71.45
		0.4	70.70	71.60	70.55	72.05	73.75	71.70
		0.6	68.60	72.65	67.95	72.60	78.40	72.15
		1.0	69.15	86.60	68.20	73.50	91.00	72.20
	negative	0.0	68.30	67.45	68.30	70.05	70.15	71.00
		0.2	68.95	69.35	68.85	70.60	72.10	70.60
		0.4	69.40	70.20	69.10	68.65	73.05	68.35
		0.6	69.05	73.80	68.65	68.00	76.75	68.25
		1.0	67.30	84.45	66.00	67.05	88.95	65.70
Positively correlated group sizes and variances (u = 3.0)								
Equi-distant	positive	0.0	61.50	61.05	61.80	63.35	67.10	61.90
		0.2	57.50	61.45	62.40	61.20	70.30	65.55
		0.4	53.25	69.85	64.15	55.15	72.25	64.70
		0.6	48.20	83.05	63.90	47.50	82.40	65.10
		1.0	39.45	85.45	62.95	37.70	83.90	64.90
	negative	0.0	62.20	59.70	60.65	62.65	61.55	64.10
		0.2	57.45	63.05	62.50	59.15	64.95	64.85
		0.4	51.20	68.90	62.75	53.40	72.75	64.80
		0.6	46.10	82.50	64.45	48.70	85.05	64.90
		1.0	38.70	85.50	62.40	44.65	88.05	62.95
Dichot-omized	positive	0.0	65.50	63.15	65.30	64.50	66.90	62.85
		0.2	60.60	66.60	67.00	62.65	69.05	66.00
		0.4	54.65	69.35	65.80	55.80	72.35	67.85
		0.6	49.25	77.40	67.30	51.10	79.85	69.70
		1.0	42.50	87.25	68.85	43.75	90.65	72.00
	negative	0.0	64.50	62.45	63.75	63.15	63.10	63.75
		0.2	59.35	65.05	65.50	59.80	67.40	66.35
		0.4	53.80	68.25	64.85	56.20	69.85	66.55
		0.6	47.90	76.05	65.95	53.15	79.65	68.50
		1.0	43.00	86.70	69.80	46.35	85.90	67.85
Negatively correlated group sizes and variances (u = 3.0)								
Equi-distant	positive	0.0	62.20	59.70	60.65	62.65	61.55	64.10
		0.2	67.15	61.40	60.15	67.15	58.25	61.40
		0.4	71.35	63.00	57.50	72.20	60.35	58.75
		0.6	76.70	72.75	57.90	79.90	71.25	59.40
		1.0	77.40	80.55	48.60	85.70	81.80	53.90
	negative	0.0	61.50	61.05	61.80	63.35	67.10	61.90
		0.2	66.60	61.00	60.05	70.00	70.05	62.75
		0.4	70.95	63.60	57.55	70.65	71.35	58.45
		0.6	76.30	70.95	55.70	75.25	79.45	57.85
		1.0	79.85	82.05	49.40	75.80	85.10	53.05
Dichot-omized	positive	0.0	64.50	62.45	63.75	63.15	63.10	63.75
		0.2	68.60	59.45	60.70	70.00	63.70	64.60
		0.4	71.85	59.25	60.85	74.45	60.70	61.90
		0.6	75.35	59.45	57.55	75.70	59.45	58.20
		1.0	81.90	75.25	51.30	86.70	82.15	56.85
	negative	0.0	65.50	63.15	65.30	64.50	66.90	62.85
		0.2	67.25	60.05	61.05	68.65	66.05	62.70
		0.4	71.10	58.40	60.35	73.60	66.10	60.80
		0.6	74.35	58.50	58.20	76.15	68.50	60.70
		1.0	82.70	76.05	54.30	81.70	78.60	56.75

pairing had been rather small for W and F* they were much more substantial at the large sample size, especially for F*. Interestingly, the interaction between the effects of variance heterogeneity and group size-variance pairing was reversed for W and F* compared to F, i.e., power was greater for positively correlated group sizes and variances for W and F* as opposed to the conservative bias introduced by this combination on F. The effect of group size inequality on this interaction at the large sample size was about the same for F, but was now also evident for the other two tests. Comparison of Figures 2 and 7 illustrates the differences between the results for small and large sample sizes when the normal population was sampled.

Several differences between the large and small sample size were observed in the results from the chi-square population. When group sizes were equal, the interaction of variance heterogeneity and mean-variance pairing was greatly diminished in the large sample size as compared to the small sample size, for all three tests. Not only was the interaction diminished in size, but the direction was reversed, so that, for the large sample size the power of the tests was greater when means and variances were positively correlated (see Figures 3 and 8).

When both group sizes and variances were unequal, the most dramatic effect of increasing total sample size on the F test was the appearance of a three-way interaction between variance heterogeneity, mean variance pairing and group size-variance pairing. Unlike the situation for the small sample size, the relationship between variance heterogeneity and mean-variance pairing, in their effects on power, was

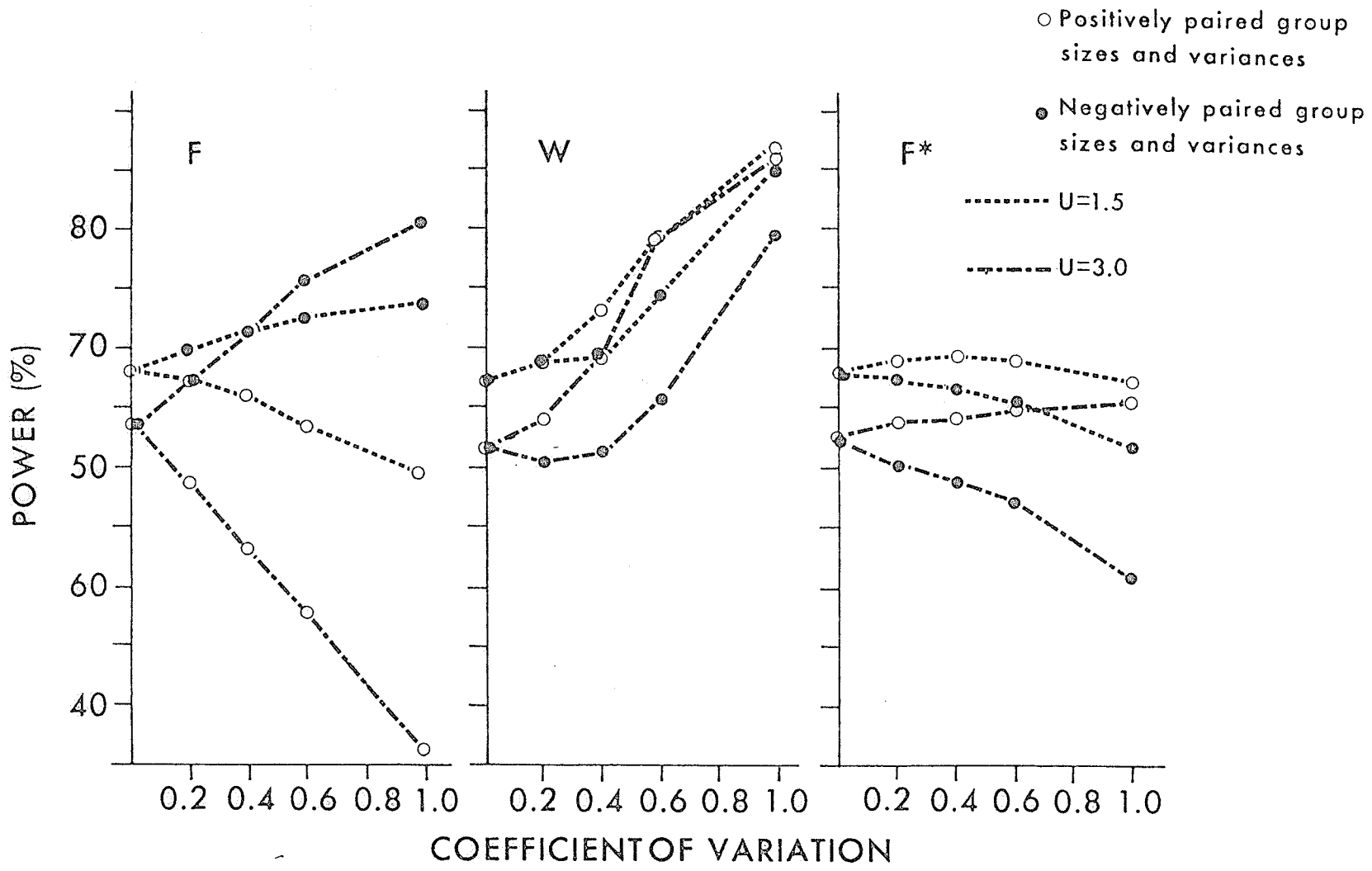


Figure 7. Power of the means tests as a function of variance heterogeneity, group size - variance pairing and group size inequality. (Normal distribution - large N - data points averaged over pattern of mean differences and mean variance pairing.)

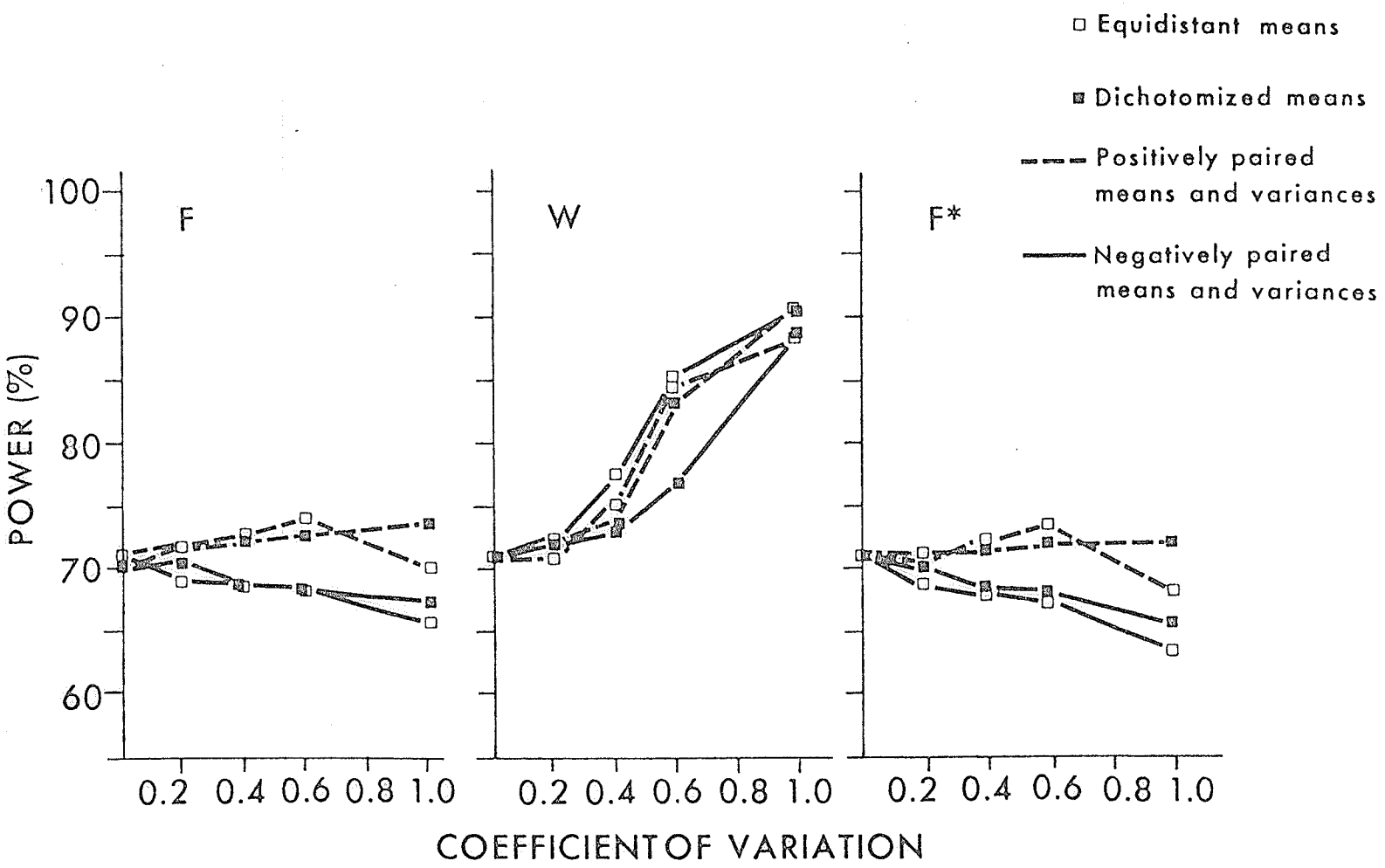


Figure 8. Power of the means tests as a function of variance heterogeneity, pattern of mean differences and mean variance pairing. (Chi-square distribution - large N - equal group sizes.)

different for the two combinations of group sizes and variances. When means and variances were positively paired, as opposed to negatively paired, there was both a greater increase in power with increasing variance heterogeneity for negatively paired group sizes and variances and a greater decrease in power with increasing c for positively paired group sizes and variances (see Figure 9 and compare with Figure 4). This was true for all levels of group size inequality, the effect not being substantially less for the smallest degree of group size inequality. Other effects observed at the small sample size were either the same or slightly increased at the larger sample size.

For the Welch test, the large effect of pairing of means and group sizes, when variances were equal, almost disappeared at the larger sample size. Since this effect is confounded with the interaction between group size-variance and mean-variance pairing (SxV), it is not surprising that the two three-way interactions involving this SxV interaction (noted at small N) were also reduced. The SxVxC interaction was only slightly less, as may be seen by comparing Figures 4 and 9, but the substantial three-way interaction between group size inequality, group size-variance pairing and mean-variance pairing, noted at the small sample size, was hardly evident at the large sample size: comparison of Figures 6 and 10 illustrates this latter point.

For F* the effects of sample size were similar to those for W; but in addition, a noticeable effect of degree of group size inequality on the interaction of variance heterogeneity and group size-variance pairing was present at large N and not at small N. The same effects were observed in the normal distribution.

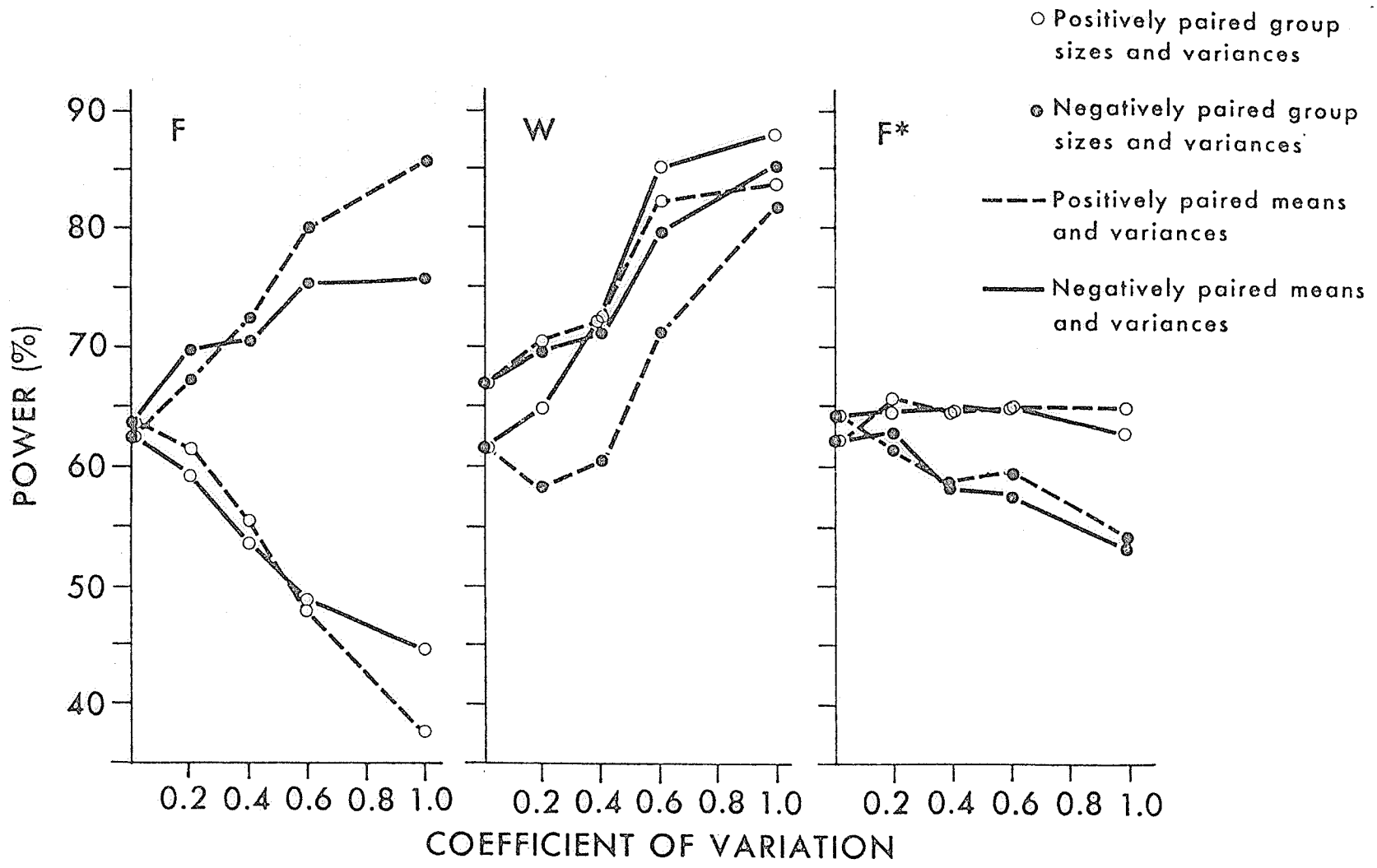


Figure 9. Power of the means tests as a function of variance heterogeneity, mean-variance pairing and group size variance pairing. (Chi-square distribution - large N - equidistant means - u = 3.0.)

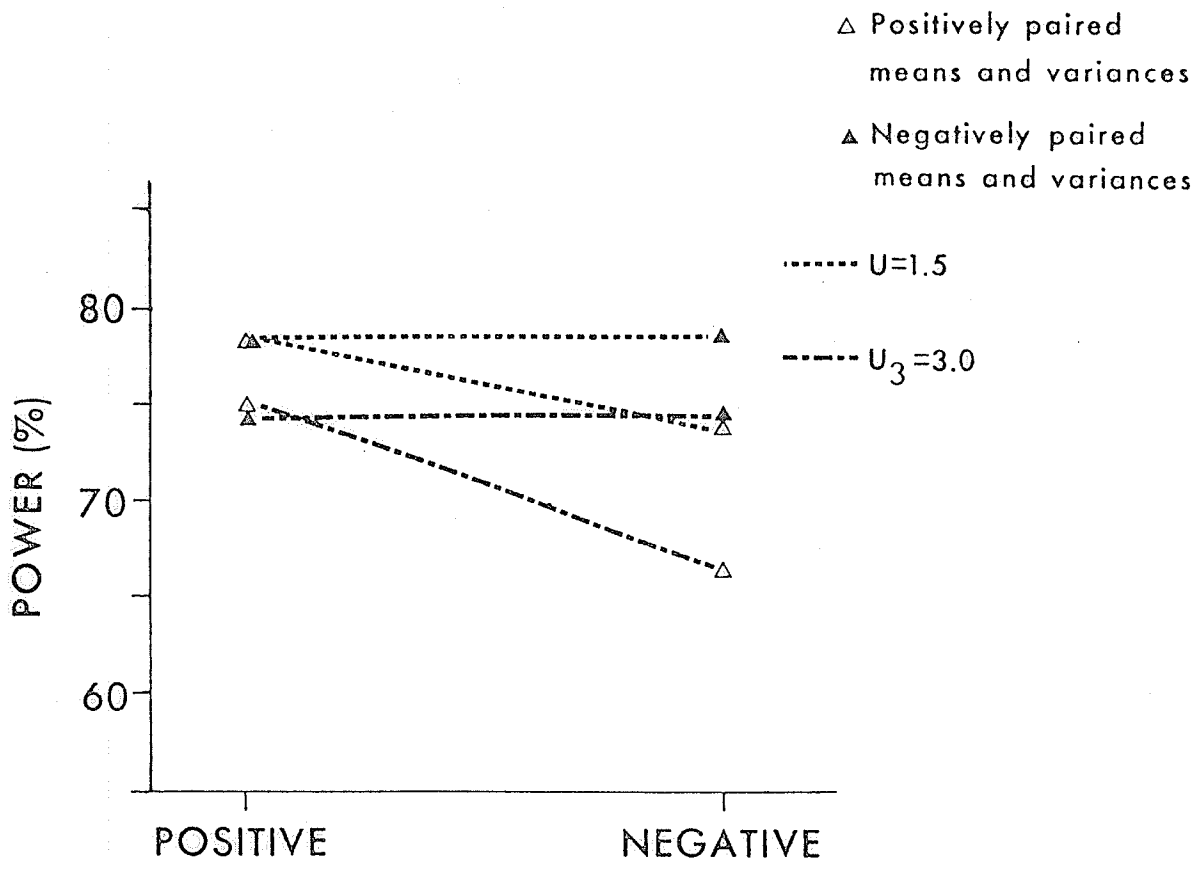


Figure 10. Power of the Welch test as a function of group size - variance pairing, mean-variance pairing and group size inequality. (Chi-square distribution - large N - equidistant means - data points averaged over degrees of variance heterogeneity.)

Summary of Results for the Means Tests

In the normal population distribution the ANOVA F-test makes too many Type I errors when a positive bias exists and too many Type II errors when a negative bias exists. On the other hand, the data which have been presented here indicate that the Welch test controls both these error rates exceptionally well, i.e., the Type I error rate is always very close to the nominal α value and the power is very close to and often exceeds a priori power calculated for the ANOVA F-test. Only very occasionally did the ANOVA F-test show adequate control of Type I error and superior empirical power to the Welch test (e.g., at $c = 0$ and $c = .2$ when group sizes and variances were negatively paired): however, this power superiority was slight ($\sim$ 5%). The Brown and Forsythe F* test controlled neither Type I nor Type II errors as well as the Welch test.

In the chi-square population the situation was not as simply defined. In the equal group size situation there is little to choose between the control of Type I error by the three tests, therefore it seems reasonable to prefer the test which has the greatest empirical power under these conditions, i.e., the Welch test. When group sizes and variances are positively correlated, the control of Type II error by the ANOVA F-test is unacceptable, and, under these circumstances, since there is again little to choose between the control of Type I error by the Welch or F* tests, the preference is for the test with the greatest power, i.e., the Welch test. When group sizes and variances are negatively correlated, the Type I error rates for the ANOVA F-test and Welch test become unacceptable, except for the ANOVA at small degrees

of variance heterogeneity (i.e., $c \leq .2$). With the exception of the highest degree of variance heterogeneity, the F* test controls Type I error rates fairly well under conditions that positively bias the ANOVA F test: additionally its empirical power is relatively close to a priori power calculated for the ANOVA F test, except for the small sample size, when means and variances are positively correlated. For this latter situation there is a case for using the ANOVA F test, when variance heterogeneity is low (i.e., $c \leq .2$), since its power is superior.

Type I Errors of the Variance Tests

Type I error rates for the variance tests are presented in Table 10. Since neither mean differences nor group size inequality produced any effects on Type I error rates, these values were obtained by averaging over dichotomized, equidistant and equal means and also over equal and unequal group sizes, so that each value was obtained from 24,000 simulations.

Table 10

Empirical Type I Error Rates (%) for the Variance Tests[a]

Population Shape	Sample Size	Box-Scheffé	Jacknife	Brown & Forsythe	Bartlett	Box-Andersen	Combined M
Normal	small	4.646	5.108	2.912	4.808	10.858	3.392
	large	4.429	5.017	4.037	4.904	6.404	3.937
Chi-square	small	4.887	12.275	4.221	42.662	11.612	9.621
	large	4.792	9.767	4.592	51.429	6.254	6.237

[a]Averaged over mean differences and group size inequality, i.e., each value was obtained from 24,000 simulations. $2\sigma_p = .0028$

In the normal population, the jacknife and Bartlett tests accurately controlled Type I errors regardless of sample size, whereas the Box-Scheffé, Brown and Forsythe, and combined M tests were conservative with the latter two being more so at the small sample size. However the Box-Andersen test was liberal and especially so at small N.

In the chi-square population, only the Box-Scheffé and Brown and Forsythe tests continued to control Type I error rates, with the latter test still being conservative. As has been observed many times before, the Bartlett test showed a huge increase in Type I error rates, up to 40-50%. The jacknife, combined M, and Box-Andersen tests showed similar degrees of non-robustness, Type I error rates being larger at the smaller sample size. However, the Type I error rates for the Box-Andersen test were approximately the same as in the normal population.

Only the Box-Scheffé and Brown and Forsythe tests were robust both to non-normality and at the small sample size. The jacknife, Bartlett, and combined M tests were robust at the small sample size, but were not robust to non-normality while the Box-Andersen test was robust to non-normality but had inflated Type I error rates at small sample size.

It can be seen that the combined M test (which is a combination of, (a) the Box-Andersen test if its correction factor, c_2, is greater than zero; and, (b) Bartlett's test if this factor is less than zero) had lower Type I error rates than the lowest of the two values of its component tests. This is because there must have been occasions on which the correction factor would have been between zero and minus two

and therefore could have turned a non-significant Bartlett value to a significant Box-Andersen value.

Power of the Variance Tests

Empirical power values for the variance tests at equal and extremely unequal group sizes (U=3.0) are presented in Tables 11 and 12 for the small and large sample sizes, respectively. In the normal population, of the tests which were robust at small sample size (i.e., all except the Box-Andersen test), the power of the tests at equal group sizes were ranked in the following decreasing order: Bartlett, jacknife, combined M, Brown and Forsythe, and Box-Scheffé. If the Box-Andersen test had been included in the above ranking, it would have appeared at different places depending on the actual degree of variance heterogeneity. This is because its power is too high at low c values, as a result of the inflated risk of Type I error, and, because its power drops at high c values for reasons discussed under Methods. When power was averaged over c values the Box-Andersen test had exactly the same power as the jacknife test. At large sample size the same rank order of tests prevailed, except that the power of the Box-Andersen test now fell between that of the jacknife and Bartlett tests.

Introduction of unequal group sizes maintained the same rank order of power of the tests with respect to each other. However, power was greater when group sizes and variances were negatively correlated as opposed to positively correlated and this effect was especially noticeable at $c = .4$ and $c = .6$, i.e., in the middle range of variance

Table 11.

Empirical power values (%) for the variance tests (small N, equal means)

Pairing of group sizes & variances	Coefficient of variation	Statistics					
		Box-Scheffe	Jacknife	Brown & Forsythe	Bartlett	Box-Andersen	Combined M
Normal Distribution							
Equal n's	0.2	7.15	9.15	6.25	10.45	16.95	7.65
	0.4	18.25	28.50	20.05	34.20	38.35	26.15
	0.6	51.90	81.45	59.85	89.80	79.90	78.10
	1.0	71.10	93.70	87.90	98.05	77.60	90.45
Positive (u = 3.0)	0.2	6.55	9.45	4.85	8.65	14.80	6.45
	0.4	12.65	20.70	11.95	22.10	28.25	16.70
	0.6	33.45	59.30	33.90	63.10	59.75	50.65
	1.0	58.00	88.25	82.35	96.80	79.50	88.05
Negative (u = 3.0)	0.2	5.85	9.20	7.00	10.05	17.55	7.00
	0.4	14.65	29.65	21.35	34.70	36.10	24.90
	0.6	51.05	84.60	69.75	91.60	70.50	78.15
	1.0	56.75	89.80	81.85	94.65	55.25	81.20
Chi-square Distribution							
Equal n's	0.2	5.00	13.55	5.80	48.30	14.45	12 40
	0.4	11.70	22.95	12.30	60.30	21.80	20.00
	0.6	27.40	46.60	31.45	87.95	44.50	44.75
	1.0	40.50	60.95	57.15	94.60	53.80	59.65
Positive (u = 3.0)	0.2	5.95	14.15	4.60	42.75	11.00	9.35
	0.4	10.40	20.80	7.00	57.60	15.95	14.30
	0.6	20.00	37.10	14 75	78.10	26.40	26.05
	1.0	37.60	56.25	41.80	98.25	44.30	45.85
Negative (u = 3.0)	0.2	5.05	14.85	6.40	44.60	13.95	11.75
	0.4	10.15	23.50	16.00	57.10	21.65	21.45
	0.6	30.95	48.90	41.15	84.75	43.05	48.55
	1.0	34.80	53.85	56.55	89.10	43.10	55.90

Table 12.

Empirical power values (%) for the variance tests (large N, equal means)

Pairing of group sizes & variances	Coeffic-ient of variation	Statistics					
		Box-Scheffe	Jacknife	Brown & Forsythe	Bartlett	Box-Andersen	Combined M
Normal Distribution							
Equal n's	0.2	17.20	24.15	19.45	26.15	28.15	22.35
	0.4	65.70	84.20	76.15	87.30	86.75	83.70
	0.6	99.55	99.95	100.00	100.00	99.90	100.00
	1.0	99.95	100.00	100.00	100.00	98.10	100.00
Positive (u = 3.0)	0.2	14.50	21.20	16.90	20.80	23.45	18.00
	0.4	49.15	72.80	60.90	75.30	73.30	69.75
	0.6	96.30	99.85	99.30	100.00	99.80	99.80
	1.0	99.95	100.00	100.00	100.00	99.20	100.00
Negative (u = 3.0)	0.2	12.05	22.70	18.35	24.60	27.15	20.25
	0.4	58.15	84.60	76.15	87.35	84.90	82.60
	0.6	99.05	100.00	99.95	100.00	99.80	100.00
	1.0	99.80	100.00	100.00	100.00	91.10	100.00
Chi-square Distribution							
Equal n's	0.2	9.35	16.10	10.75	61.80	11.55	11.55
	0.4	31.45	40.40	38.85	87.70	34.95	34.95
	0.6	85.66	85.70	91.90	99.80	80.60	80.60
	1.0	96.35	93.35	99.70	100.00	92.75	92.90
Positive (u = 3.0)	0.2	8.85	15.60	8.10	60.80	9.20	9.20
	0.4	26.00	34.65	24.75	82.85	23.05	23.05
	0.6	70.30	74.35	68.70	98.35	57.85	57.85
	1.0	94.20	92.95	98.60	99.90	86.30	86.35
Negative (u = 3.0)	0.2	7.70	16.70	11.75	59.90	12.55	12.55
	0.4	60.30	39.70	44.95	83.90	39.50	39.50
	0.6	85.70	86.85	95.45	99.65	88.30	88.50
	1.0	91.80	89.95	98.45	99.95	91.85	93.00

heterogeneity. Power of the tests, when group sizes and variances were negatively correlated, approximated that of the equal group size condition. This relationship of power to pairings of group sizes and variances was equally valid at both the large and small sample size. Increasing group size inequality accentuated the effect of power reduction in the middle variance heterogeneity range for positively correlated group sizes and variances, therefore Tables 11 and 12 illustrate the extremes of this effect.

In the chi-square population for the small sample size condition, comparison of the behaviour of the two tests which control Type I error, namely the Box-Scheffé and Brown and Forsythe tests, reveals that the power preference depends upon the relationship between group sizes and variances. If n_j's are equal, or, unequal and negatively correlated with variances, the Brown and Forsythe test has superior power, whereas, if larger groups have larger variances, the Box-Scheffé test has greater power. The Box-Andersen and combined M tests, which have empirical alpha values of about twice the nominal level, demonstrate approximately equal power values, that are somewhat higher than those of the two robust tests. The jacknife and Bartlett tests have the second highest and highest power, respectively, however the former has a Type I error rate greater than twice nominal alpha and the latter is completely unacceptable because of its 50% Type I error rates.

At the large sample size, in the chi-square population, the Type I error rates of the Box-Scheffé, Brown and Forsythe, Box-Andersen, and combined M tests are all acceptable. The same power relationships

exist between the former two tests, with respect to the group size-variance relationship, as occurred at small N. Power of the Box-Andersen and combined M tests is identical, falling between that of the Box-Scheffé and Brown and Forsythe tests for equal n_j's and negatively paired n_j's and variances, and, is less than either for positively paired n_j's and variances. The jacknife test, which has an empirical alpha of twice the nominal value, is more powerful when n_j's are equal or unequal and positively correlated with variances, than the four tests which are robust at large sample size, but has lower power than the Brown and Forsythe test at $c \geq .4$ for negatively paired group sizes and variances.

Summary of Results for the Variance Tests

Overall, the preferred variance test seems to be the Brown and Forsythe test on absolute deviations from the median. Not only does this test control Type I error rates but it is also more powerful than the other truly robust test, i.e., the Box-Scheffé. Although there are specific situations when the Box-Scheffé test is actually more powerful, the excess is not great, and averaged over all conditions (where the group means are equal) the power of the Box-Scheffé and Brown and Forsythe tests is, respectively, 42.90 and 48.82. Since these tests have rather low power at the low end of the variance heterogeneity continuum, it may be preferable, at times, to sacrifice control of Type I error and use a less robust test, in order to gain more power to detect a small difference in variances. This, of course, depends on the relative

cost, in a given situation, of making a Type I as opposed to a Type II error, a matter of some concern when performing variance tests as a preliminary to a test of mean differences.

Type I Errors of the Sequential Testing Procedures

Type I error rates for all sequential procedures and individual means tests, under all conditions when group sizes are equal, are presented in Table 13. In general, for the sequential procedures, Type I error rates (and power) were closer to the ANOVA values when variance heterogeneity was small, and closer to the values for the alternate test (i.e., Welch or Brown and Forsythe's F*) when variance heterogeneity was large. This relationship exists because, when the power of any variance test is less than 50%, the ANOVA F-test on means will be performed more often than the alternate test; and, conversely, when the power of the variance test is greater than 50%, the alternate test will be performed more often. Naturally, the power of any variance test is lower at low degrees of variance heterogeneity and is therefore more likely to be less than 50%. Also, therefore, Type I error rates (and power) of the sequential procedures approximated those of the alternate test more often at the larger sample size because the variance tests are more powerful at large N.

When sampling from the normal population using equal group sizes, it is clear that none of the sequential procedures, for either alternate test, showed better control of Type I errors than did the Welch test alone. This is inevitably so, since the performance of the Welch test

Table 13.

Type I Error Rates (%) for the Individual Means Tests and Sequential Procedures.
(Equal n_j's)

Coefficient of variation of the variances c	Combined ANOVA and Welch Procedures								Combined ANOVA and F* Procedures							
	ANOVA	WELCH	FW/BS	FW/JK	FW/BF	FW/BA	FW/B	FW/CM	ANOVA	F*	FF*/BS	FF*/JK	FF*/BF	FF*/BA	FF*/B	FF*/CM
Normal Distribution (Small N)																
0.0	5.25	5.05	5.45	5.45	5.35	5.60	5.30	5.35	5.25	5.05	5.20	5.20	5.20	5.20	5.15	5.20
0.2	5.15	5.30	5.15	5.55	5.45	5.55	5.55	5.40	5.15	4.90	5.10	5.15	5.15	5.10	5.10	5.15
0.4	6.15*	4.75	6.15*	6.25*	6.55*	6.20*	6.25*	6.35*	6.15*	5.70	6.00*	5.95	6.05*	5.85	5.95	6.00*
0.6	5.85	5.55	6.80*	6.40*	6.70*	6.15*	6.20*	6.50*	5.85	5.40	5.65	5.40	5.45	5.40	5.40	5.40
1.0	8.50*	5.25	7.00*	5.75	6.15*	5.85	5.40	5.80	8.50*	6.85*	7.50*	7.00*	7.20*	7.35*	6.85*	7.05*
Normal Distribution (Large N)																
0.0	4.80	5.05	4.85	5.00	4.95	5.05	5.05	5.00	4.80	4.80	4.80	4.80	4.80	4.80	4.80	4.80
0.2	5.90	5.95	5.95	6.05*	6.00*	5.95	6.05*	6.05*	5.90	5.90	5.90	5.90	5.90	5.90	5.90	5.90
0.4	5.80	5.45	5.75	5.65	5.80	5.70	5.60	5.70	5.80	5.65	5.70	5.70	5.75	5.70	5.70	5.70
0.6	6.35*	5.50	5.55	5.55	5.50	5.50	5.50	5.50	6.35*	6.30*	6.30*	6.30*	6.30*	6.30*	6.30*	6.30*
1.0	7.75*	4.80	4.80	4.80	4.80	4.70	4.80	4.80	7.75*	7.35*	7.35*	7.35*	7.35*	7.35*	7.35*	7.35*
Chi-Square Distribution (Small N)																
0.0	3.80**	4.45	4.35	4.85	3.10**	3.65**	4.95	3.75**	3.80**	3.25**	3.65**	3.60**	3.55**	3.60**	3.35**	3.55**
0.2	4.50	5.60	4.75	5.00	4.10	4.70	5.85	4.60	4.50	3.95**	4.45	4.20	4.30	4.30	4.15	4.30
0.4	5.25	6.30*	5.40	5.80	4.55	5.10	6.45*	5.10	5.20	4.30	4.85	4.70	4.60	4.65	4.40	4.60
0.6	6.25*	9.20*	7.20*	7.95*	7.60*	7.60*	9.15*	7.70*	6.25*	5.30	5.95	5.45	5.55	5.55	5.40	5.60
1.0	9.55*	8.85*	9.65*	9.95*	9.90*	9.30*	9.45*	9.15*	9.55*	8.30*	8.85*	8.55*	8.55*	8.70*	8.30*	8.50*
Chi-Square Distribution (Large N)																
0.0	5.00	5.85	5.20	5.85	4.75	5.35	5.90	5.35	5.00	4.75	5.00	4.85	4.95	4.90	4.75	4.90
0.2	4.90	6.25*	5.25	5.50	5.00	5.20	6.00*	5.20	4.90	4.75	4.90	4.90	4.80	4.90	4.75	4.90
0.4	6.25*	6.00*	5.80	6.10*	5.60	5.60	6.05*	5.60	6.25*	6.15*	6.25*	6.15*	6.15*	6.15*	6.15*	6.15*
0.6	5.80	5.90	5.45	5.65	5.65	5.35	5.90	5.35	5.80	5.55	5.55	5.55	5.55	5.55	5.55	5.55
1.0	7.95*	6.45*	6.85*	6.95*	6.50*	6.70*	6.45*	6.60*	7.95*	7.75*	7.75*	7.75*	7.75*	7.75*	7.75*	7.75*

* Type I error rates $2\sigma_p$ greater than α, where $\sigma_p = (\alpha(1 - \alpha)/2{,}000)^{1/2} = .0097$

** Type I error rates $2\sigma_p$ less than α.

is the best of all the tests of mean differences. In order for any sequential procedure to be as good as the Welch test, the variance test involved would have to have a power of 100% at all levels of variance heterogeneity -- an obvious impossibility. Of the sequential procedures involving the Welch test as the alternate test to the ANOVA when the assumption of homogeneity of variance was rejected, that using Bartlett's variance test performed best; the same was true when using F* as the alternate means test. The foregoing held true for both small and large N. (When N=144, F and F* behave almost identically; and, therefore, all sequential tests involving F* as the alternate are also almost identical to both F and F*.)

For equal group sizes sampled from the Chi-square distribution, the situation changes considerably. When the sample size is small, the Welch test and the sequential procedures in which it is the alternate test, perform worse overall than the ANOVA F-test. The best of the combined ANOVA F-test and Welch procedures is that using the Box-Scheffé variance test. In contrast, the sequential procedures combining F and F* showed smaller deviations from nominal α , overall, than did either test separately, the best of these being that using the Jacknife variance test. When the sample size is large, the situation improves for the Welch sequential procedures because the Welch test becomes more robust. In fact, all sequential procedures for either of the alternate means test are superior to either of their component means tests. For the Welch and ANOVA F-test combined procedures, that using the Brown and Forsythe variance test was best, whereas there was little to choose

between the sequential procedures combining the ANOVA F and F* tests.

When group sizes are unequal, different situations obtain depending upon whether group sizes and variances are positively or negatively paired. These results are presented in Table 14. When small samples were taken from the normal population and group sizes and variances were positively paired, the situation is analogous to the equal n_j condition in that the Welch test is superior to all single and combined tests. In general, the sequential procedures err on the conservative side because (a) the ANOVA F-test is conservative under these conditions, (b) the power of the variance tests is at its lowest with this pairing of n_j's and variances, and therefore (c) the ANOVA F-test will be chosen more frequently because of failure to reject the homogeneity of variance assumption. Of the sequential procedures combining the ANOVA F-test and the Welch test, those using the combined M and Box-Andersen variance tests were best; whereas, for the sequential procedures combining the F and F* tests, the best procedures were those using the Box-Scheffé and Box-Andersen variance tests. These were also an improvement on uniformly adopting the F* test. At large sample size, preferences were essentially in the same order although none of the sequential tests was now significantly conservative.

When small samples were taken from the chi-square population and group sizes and variances were positively paired, the Welch test showed the largest deviations from nominal α , as when group sizes were equal. However, unlike the latter situation, the sequential procedures involving the Welch test were now an improvement on the use of either

Table 14.

Type I Error Rates (%) for the Individual Means Tests and Sequential Procedures
(Unequal n_j's (u=3.0) positively paired with variances)

Coefficient of variation of the variances c	Combined ANOVA and Welch Procedures								Combined ANOVA and F* Procedures							
	ANOVA	WELCH	FW/BS	FW/JK	FW/BF	FW/BA	FW/S	FW/CM	ANOVA	F*	FF*/BS	FF*/JK	FF*/BF	FF*/BA	FF*/B	FF*/CM
Normal Distribution (Small N)																
0.0	5.00	5.45	5.40	5.50	5.40	5.75	5.55	5.55	5.00	4.65	5.15	5.05	5.05	5.25	5.10	5.15
0.2	4.20	5.60	4.65	5.30	4.70	5.35	5.20	5.00	4.20	5.30	4.30	4.55	4.25	4.40	4.45	4.40
0.4	2.80**	4.70	3.70**	3.90**	3.60**	3.90**	4.15	3.85**	2.80**	5.50	3.20**	3.35**	3.35**	3.50**	3.60**	3.40**
0.6	2.85**	4.90	4.15	4.60	4.45	4.80	5.20	4.80	2.85**	5.95	3.85**	4.65	4.00**	4.50	4.95	4.35
1.0	3.25**	5.75	5.40	5.60	5.65	5.05	5.35	5.55	3.25**	7.55*	5.50	6.90*	6.65*	6.45*	7.50*	7.00*
Normal Distribution (Large N)																
0.0	4.75	5.05	4.85	4.90	4.85	4.85	4.90	4.80	4.75	4.90	4.80	4.85	4.80	4.80	4.75	4.75
0.2	3.90**	5.15	4.30	4.50	4.35	4.50	4.45	4.45	3.90**	5.40	4.20	4.45	4.30	4.40	4.40	4.40
0.4	3.10**	5.20	4.25	4.85	4.70	4.90	4.95	4.85	3.10**	5.85	4.45	4.80	4.85	5.00	5.00	4.90
0.6	2.90**	4.90	4.90	4.90	4.90	4.90	4.90	4.90	2.90**	6.15*	6.00*	6.15*	6.15*	6.15*	6.15*	6.15*
1.0	3.10**	5.90	5.90	5.90	5.90	5.90	5.90	5.90	3.10**	7.75*	7.75*	7.75*	7.75*	7.70*	7.75*	7.75*
Chi-Square Distribution (Small N)																
0.0	4.55	7.65*	4.65	6.20*	3.55**	5.25	7.90*	5.00	4.55	3.95**	4.35	4.20	3.75**	4.60	4.15	4.25
0.2	3.95**	6.80*	5.25	5.75	3.85**	4.60	6.95*	4.70	3.95**	4.45	4.50	4.25	4.35	4.00**	4.60	4.10
0.4	3.35**	5.30	4.70	5.15	3.75**	4.75	5.75	4.90	3.35**	4.55	3.75**	3.85**	4.10	4.10	4.75	4.25
0.6	3.65**	7.75*	5.20	6.90*	5.00	5.50	7.75*	5.45	3.65**	5.70	4.30	4.70	4.70	4.70	5.50	4.75
1.0	4.25	6.70*	6.25*	6.85*	6.30*	6.10*	7.00*	6.25*	4.25	7.80*	5.75	6.40*	5.65	6.00*	7.50*	6.05*
Chi-Square Distribution (Large N)																
0.0	5.30	7.05*	5.70	5.85	4.60	5.00	7.00*	5.00	5.30	4.70	5.10	5.05	4.75	4.85	4.70	4.85
0.2	4.00**	6.20*	5.80	6.00*	5.05	5.55	6.50*	5.55	4.00**	5.75	5.10	5.35	5.50	5.15	5.85	5.15
0.4	2.50**	5.10	4.65	4.70	4.50	4.45	5.25	4.45	2.50**	5.15	4.50	4.30	4.50	4.20	5.15	4.20
0.6	2.65**	5.30	4.70	5.00	4.80	4.40	5.25	4.40	2.65**	5.85	5.35	5.40	5.45	5.00	5.85	5.00
1.0	3.40**	6.25*	6.35*	6.40*	6.30*	6.25*	6.25*	6.25*	3.40**	8.15*	7.60*	7.85*	7.90*	7.45*	8.15*	7.45*

* Type I error rates $2\sigma_p$ greater than α, where $\sigma_p = (\alpha(1 - \alpha)/2{,}000)^{\frac{1}{2}} = .0097$.

** Type I error rates $2\sigma_p$ less than α.

Table 14.(continued)

Type I Error Rates (%) for the Individual Means Tests and Sequential Procedures.
(Unequal n_j's (u=3.0) negatively paired with variances)

Coefficient of variation of the variances	Combined ANOVA and Welch Procedures.								Combined ANOVA and F* Procedures.							
c	ANOVA	WELCH	FW/BS	FW/JK	FW/BF	FW/BA	FW/B	FW/CM	ANOVA	F*	FF*/BS	FF*/JK	FF*/BF	FF*/BA	FF*/B	FF*/CM
Normal Distribution (Small N)																
0.0	5.00	5.45	5.40	5.50	5.40	5.75	5.55	5.55	5.00	4.65	5.15	5.05	5.05	5.25	5.10	5.15
0.2	5.75	4.25	5.90	6.00*	5.95	5.60	5.95	5.95	5.75	3.75**	5.70	5.40	5.50	5.10	5.45	5.60
0.4	9.60*	5.80	9.05*	8.40*	8.40*	7.95*	8.40*	8.40*	9.60*	4.85	8.60*	7.95*	7.95*	7.45*	7.40*	7.70*
0.6	13.55*	5.25	9.70*	6.80*	8.55*	7.45*	6.45*	7.35*	13.55*	5.65	10.15*	6.90*	8.50*	7.90*	6.35*	7.90*
1.0	20.70*	6.40*	13.35*	8.20*	9.15*	12.85*	7.20*	9.25*	20.70*	7.25*	13.30*	8.80*	9.35*	13.35*	7.90*	9.85*
Normal Distribution (Large N)																
0.0	4.75	5.05	4.85	4.90	4.85	4.85	4.90	4.80	4.75	4.90	4.80	4.85	4.80	4.80	4.75	4.75
0.2	8.50*	5.80	8.10*	7.40*	7.75*	7.35*	7.60*	7.75*	8.50*	6.00*	8.05*	7.35*	7.70*	7.40*	7.45*	7.60*
0.4	10.05*	5.00	6.95*	5.85	6.30*	5.80	5.55	5.90	10.05*	5.60	7.40*	6.30*	6.80*	6.30*	6.05*	6.40*
0.6	13.20*	4.90	5.10	4.90	4.90	4.90	4.90	4.90	13.20*	5.85	6.00*	5.85	5.85	5.85	5.85	5.85
1.0	20.30*	5.30	5.35	5.30	5.30	6.50*	5.30	5.30	20.30*	7.70*	7.75*	7.70*	7.70*	8.80*	7.70*	7.70*
Chi-Square Distribution (Small N)																
0.0	4.55	7.65*	4.65	6.20*	3.55**	5.25	7.90*	5.00	4.55	3.95**	4.35	4.20	3.75**	4.60	4.15	4.25
0.2	5.50	8.15*	5.65	6.15*	4.65	5.45	7.25*	5.05	5.50	4.40	5.00	5.25	4.60	5.15	4.60	4.90
0.4	10.05*	12.30*	9.50*	10.20*	8.20*	9.25*	11.11*	8.55*	10.05*	6.35*	8.65*	8.25*	7.25*	8.05*	6.40*	7.25*
0.6	13.65*	13.40*	11.20*	12.55*	10.65*	12.55*	13 25*	11.25*	13.65*	5.95	9.30*	8.50*	7.60*	9.80*	6.55*	7.85*
1.0	22.40*	12.70*	18.10*	16.80*	16.05*	17.90*	13.75*	16.15*	22.40*	10.30*	16.90*	14.65*	14.15*	16.85*	11.10*	14.35*
Chi-Square Distribution (Large N)																
0.0	5.30	7.05*	5.70	5.85	4.60	5.00	7.00*	5.00	5.30	4.70	5.10	5.05	4.75	4.85	4.70	4.85
0.2	5.70	6.10*	5.10	5.20	4.40	4.60	5.65	4.60	5.70	4.10	4.70	4.35	4.25	4.25	3.95**	4.25
0.4	9.95*	7.20*	7.85*	7.60*	6.75*	7.00*	7.00*	7.00*	9.95*	5.70	7.75*	6.80*	6.45*	6.60*	5.95	6.60*
0.6	14.90*	10.65*	10.50*	10.45*	10.45*	10.45*	10.65*	10.25*	14.90*	7.20*	8.00*	8.00*	7.45*	7.85*	7.25*	7.70*
1.0	20.95*	10.30*	10.90*	11.55*	10.50*	11.35*	10.35*	10.90*	20.95*	9.65*	10.40*	10.75*	9.75*	10.70*	9.65*	10.35*

* Type I error rates $2\sigma_p$ greater than α, where $\sigma_p = (\alpha(1 - \alpha)/2{,}000)^{\frac{1}{2}} = .0097$.

** Type I error rates $2\sigma_p$ less than α.

the ANOVA F-test or the Welch test alone, the best being that using the combined M variance test. The combined procedures using the F* and ANOVA F-tests also showed better control of Type I error rates than each test alone, with the best combination being that using the Box-Andersen test. This sequential procedure was not as good as the best ANOVA-Welch sequential procedure. Essentially the same situation existed when using large unequal samples from the chi-square population.

When group sizes and variances are negatively paired, Type I error rates were seriously inflated for the ANOVA F-test in both populations and for the Welch test in the chi-square population. In the normal distribution, no sequential procedure improved upon the performance of the Welch test alone, especially when the sample size was large. Similarly, for the combined ANOVA F and F* procedures, none controlled Type I errors as well as the F* test alone. The best of the sequential procedures in all cases were those using Bartlett's variance test.

For negative pairing of group sizes and variances in the chi-square population, all sequential procedures combining the ANOVA F and Welch tests showed superior control of Type I error rates compared to either test individually, the best procedure being that using the combined M variance test. The relative behaviour of the tests was the same for both sample sizes, but control of Type I error rates was better at large N. Combined procedures using the ANOVA F and F* tests did not improve on the F* test alone at either sample size in the chi-square population, and the F* test alone was an improvement on the best sequential procedure using the combined ANOVA F and Welch tests. The

best sequential procedure combining the ANOVA F and F* tests was that using the Bartlett variance test.

Summary of Type I Error Rates of Combined Versus Individual Means Tests.

When the population sampled is normal, no sequential testing procedure is preferable to uniformly adopting the Welch test regardless of whether group sizes and/or variances are unequal. When the chi-square population is sampled and n_j's are equal, sequential procedures using F* as the alternate test are preferable for N=48, and procedures using the Welch test as the alternate test are preferable for N=144. A more critical situation exists when n_j's and variances are both unequal. If pairing of n_j's and variances is positive, sequential procedures combining the ANOVA F test and Welch test are to be preferred; but, when the pairing of n_j's and variances is negative, uniformly adopting the Brown and Forsythe F* test is best for both sample sizes.

Power of the Sequential Testing Procedures

Discussion of the power of the sequential procedures will be limited to the situations in which a sequential procedure improved on the control of Type I error rates. This only occurred when the chi-square population was sampled and was of most importance under conditions where the ANOVA F-test is notably conservative, i.e., when variances are heterogeneous and positively paired with unequal group sizes. These results are presented in Table 15. In the chi-square population, especially at small N, the Welch test significantly exceeded

Table 15.

Power (%) for the Individual Means Tests and Sequential Procedures in the Chi-Square Population.[a]
(unequal n_j's (u=3.0) positively paired with variances)

Coefficient of variation of the variances c	Combined ANOVA and Welch Procedures								Combined ANOVA and F* Procedures							
	ANOVA	WELCH	FW/BS	FW/JK	FW/BF	FW/BA	FW/B	FW/CM	ANOVA	F*	FF*/BS	FF*/JK	FF*/BF	FF*/BA	FF*/B	FF*/CM
Positively Paired Means and Variances (Small N)																
0.0	24.33	36.45	24.93	26.60	24.50	25.53	32.08	24.93	24.33	26.60	24.28	24.30	24.00	24.00	25.48	23.98
0.2	18.10	34.38	19.48	22.03	18.45	20.30	29.50	19.98	18.10	25.40	18.60	19.68	18.23	19.35	23.18	19.03
0.4	12.73	33.88	16.65	21.03	13.28	17.00	30.88	16.58	12.73	23.15	14.30	16.60	13.65	15.55	21.30	15.28
0.6	9.95	33.15	19.08	25.70	13.58	19.43	32.55	19.38	9.95	22.13	14.23	17.23	14.05	15.83	21.63	15.73
1.0	6.83	38.30	24.20	31.55	23.45	24.68	38.30	25.43	6.83	19.58	14.38	16.40	15.95	15.08	19.35	15.73
Positively Paired Means and Variances (Large N)																
0.0	63.93	67.00	63.98	64.53	64.10	64.08	66.34	64.08	63.93	62.38	63.75	63.43	63.68	63.73	62.95	63.73
0.2	61.93	69.68	62.13	62.70	61.95	62.05	67.73	62.05	61.93	65.78	62.10	62.35	61.90	62.05	64.58	62.05
0.4	55.48	72.30	58.28	60.98	56.00	57.40	71.03	57.40	55.48	66.28	57.37	59.25	56.38	56.93	65.38	56.93
0.6	49.30	81.13	72.43	73.73	68.55	66.15	81.03	66.15	49.30	67.40	62.93	63.43	62.55	59.55	67.30	59.55
1.0	40.73	87.28	84.65	84.05	86.60	78.83	87.28	78.83	40.73	68.45	67.58	66.70	68.18	64.48	68.45	64.48
Negatively Paired Means and Variances (Small N)																
0.0	25.40	19.08	25.03	24.13	24.73	24.15	23.35	24.25	25.40	19.33	24.70	23.75	24.73	24.30	21.95	24.30
0.2	23.88	22.55	23.85	24.38	23.90	23.63	25.65	24.00	23.88	22.65	23.70	23.33	23.90	23.55	23.25	23.65
0.4	21.88	29.00	22.70	24.60	22.68	23.50	28.98	23.53	21.28	24.75	21.75	22.28	21.75	22.43	23.73	22.33
0.6	18.83	41.28	24.35	30.33	23.73	25.68	38.95	25.70	18.83	26.98	20.70	22.28	20.23	21.28	25.33	21.20
1.0	20.45	45.85	32.22	38.10	35.65	33.68	45.38	34.75	20.45	30.70	23.73	25.68	24.13	24.90	30.08	25.15
Negatively Paired Means and Variances (Large N)																
0.0	62.90	62.33	62.70	62.78	62.93	62.95	62.95	62.95	62.90	63.93	62.93	62.95	62.98	63.03	63.73	63.03
0.2	59.48	66.18	60.03	61.08	60.48	60.30	65.45	60.30	59.48	65.60	60.05	60.58	60.23	60.13	64.25	60.13
0.4	54.80	71.30	60.20	63.20	60.60	59.98	70.68	59.98	54.80	65.68	58.15	59.93	58.08	58.03	65.23	58.03
0.6	50.93	82.35	75.50	76.43	76.15	71.05	82.28	71.05	50.93	66.70	63.15	63.48	63.15	61.20	66.65	61.20
1.0	45.50	86.98	86.28	85.40	86.90	81.35	86.98	81.38	45.50	65.40	64.93	64.60	65.33	62.58	65.40	62.58

a Each value in the table was based on 4,000 simulations.

Table 15. (continued)

Power (%) for the Individual Means Tests and Sequential Procedures in the Chi-Square Population[a]
(Unequal n_j's (u=3.0) negatively paired with variances)

Coefficient of variation of the variances	Combined ANOVA and Welch Procedures.								Combined ANOVA and F* Procedures.						
c	ANOVA	WELCH	FW/BS	FW/JK	FW/BF	FW/BA	FW/B	FW/CM	ANOVA	F*	FF*/BS	FF*/JK	FF*/BF	FF*/BA	FF*/B	FF*/CM
Positively Paired Means and Variances (Small N)																
0.0	25.40	19.08	25.03	24.13	24.73	24.15	23.35	24.25	25.40	19.33	24.70	23.75	24.73	24.30	21.95	24.30
0.2	27.40	16.13	26.25	24.43	24.45	24.65	20.60	24.30	27.40	16.25	26.08	23.58	24.73	24.88	19.23	24.23
0.4	28.95	13.05	25.78	23.18	21.88	22.63	16.70	21.80	28.95	12.90	25.48	21.55	22.23	22.68	15.38	21.70
0.6	34.10	12.25	24.10	20.63	17.60	21.33	13.50	17.93	34.10	12.28	23.68	19.85	17.90	21.80	13.03	18.40
1.0	43.03	20.00	31.68	29.40	24.83	31.50	20.85	25.15	43.03	13.45	27.05	23.50	19.18	28.28	14.08	19.93
Positively Paired Means and Variances (Large N)																
0.0	62.90	62.33	62.70	62.78	62.93	62.95	62.95	62.95	62.90	63.93	62.93	62.95	62.98	63.03	63.73	63.03
0.2	68.58	60.98	68.10	67.25	67.70	67.70	63.88	67.70	68.58	63.00	68.25	67.28	68.20	68.10	64.40	68.10
0.4	73.33	60.53	71.10	68.68	68.83	69.48	62.23	69.48	73.33	60.33	70.18	67.45	68.85	69.10	61.30	69.10
0.6	77.80	65.35	67.53	67.58	65.65	66.75	65.38	66.75	77.80	58.80	61.23	60.70	59.08	69.33	58.85	61.33
1.0	86.20	81.98	82.78	83.05	82.20	81.89	81.98	81.80	86.20	55.38	57.75	57.93	55.60	57.08	55.38	56.95
Negatively Paired Means and Variances (Small N)																
0.0	24.33	36.45	24.93	26.60	24.50	25.53	32.08	24.93	24.33	26.60	24.28	24.30	24.00	24.00	25.48	23.98
0.2	30.88	38.45	31.20	32.75	31.20	32.28	36.50	32.10	30.88	28.73	30.23	29.63	29.63	30.00	28.93	29.73
0.4	37.70	40.73	38.00	39.08	38.25	38.40	41.78	38.68	37.70	29.33	36.60	34.73	35.15	35.10	31.88	34.63
0.6	44.83	45.53	45.38	46.28	45.93	44.75	46.83	45.18	44.83	29.45	39.93	36.75	37.23	38.38	31.10	36.60
1.0	52.20	50.38	52.78	52.85	54.28	51.05	51.28	51.95	52.20	30.85	44.93	39.98	39.25	43.30	31.63	38.58
Negatively Paired Means and Variances (Large N)																
0.0	63.93	67.00	63.98	64.53	64.10	64.08	66.34	64.08	63.93	62.38	63.75	63.43	63.68	63.73	62.95	63.73
0.2	69.33	68.05	69.30	69.40	69.65	69.55	69.65	69.55	69.33	62.73	68.28	67.65	67.83	67.93	64.30	67.98
0.4	72.13	68.73	71.18	70.93	70.75	70.50	69.10	70.50	72.13	59.63	67.63	65.33	64.48	65.40	59.93	65.40
0.6	75.70	73.98	74.15	74.13	73.98	73.85	73.98	73.85	75.70	59.28	60.43	60.58	59.48	60.63	59.28	60.63
1.0	78.75	81.85	81.80	81.68	81.85	81.65	81.85	81.68	78.75	54.90	55.45	55.95	54.93	56.10	54.90	55.45

a Each value in the table was based on 4,000 simulations.

the nominal α values, while the ANOVA F-test had empirical α values significantly less than nominal α. At small N, the sequential procedures choosing between these two tests, that show best control of Type I error rates are, in order, those using the combined M, Box-Scheffé and Box-Andersen variance tests; at large N the order of preference is the Brown and Forsythe followed by either the combined M or the Box-Andersen test. Since there is little to recommend one over the other of the three best sequential procedures at either sample size in terms of control of Type I error, it seems reasonable to prefer the procedure having most power, which was that using the Box-Andersen variance test at small N and that using the Brown and Forsythe test at large N. However, in the interests of consistency, since there was little to choose between the two procedures at large N, the Box-Andersen test is recommended for selecting between the ANOVA F-test and the Welch test when group sizes and variances are positively paired.

When n_j's were equal and small, the recommended sequential procedure for controlling Type I error (i.e., that combining the ANOVA F and F* tests and using the jacknife variance test to decide between them) was also more powerful than the F* test alone but was not as powerful as the ANOVA F-test alone. The recommended procedure, when N was large (i.e., the Brown and Forsythe variance test choosing between the ANOVA F and Welch tests), was almost as powerful as the Welch test alone and was therefore a considerable improvement on the ANOVA F-test (see Figure 16).

Table 16.

Power (%) for the Individual Means Tests and Sequential Procedures in the Chi-Square Population[a]

(Equal n_j's)

Coefficient of variation of the variances	Combined ANOVA and Welch Procedures.								Combined ANOVA and F* Procedures.							
c	ANOVA	WELCH	FW/BS	FW/JK	FW/BF	FW/BA	FW/B	FW/CM	ANOVA	F*	FF*/BS	FF*/JK	FF*/BF	FF*/BA	FF*/B	FF*/CM
Positively Paired Means and Variances (Small N)																
0.0	28.38	30.78	28.45	29.18	28.03]	28.25	31.40	28.23	28.38	26.58	28.20	28.03	28.20	28.05	27.23	28.10
0.2	26.13	28.00	26.35	27.28	25.53	25.88	28.30	25.90	26.13	24.63	26.03	25.68	26.03	25.78	24.85	25.83
0.4	24.13	26.83	24.83	25.95	22.78	24.08	27.35	23.98	24.13	22.18	23.48	22.98	23.60	23.25	22.35	23.23
0.6	21.98	22.33	22.70	23.60	20.20	21.68	22.55	21.45	21.98	18.95	20.40	19.73	20.13	20.13	18.98	20.03
1.0	20.85	30.38	26.75	29.28	26.35	26.88	30.63	26.70	20.85	17.00	18.33	17.73	17.53	18.08	17.05	17.60
Positively Paired Means and Variances (Large N)																
0.0	70.58	70.60	70.70	70.78	70.65	70.60	71.28	70.60	70.58	70.50	70.55	70.53	70.55	70.58	70.50	70.58
0.2	71.48	71.68	71.53	71.58	71.53	71.48	72.00	71.48	71.48	71.33	71.48	71.43	71.48	71.48	71.33	71.48
0.4	72.28	74.45	72.73	73.38	72.10	72.55	74.70	72.55	72.28	71.98	72.18	72.15	72.25	72.18	72.00	72.18
0.6	73.28	81.45	80.95	81.03	80.73	79.68	81.45	79.68	73.28	72.73	72.83	72.75	72.75	72.80	72.73	72.80
1.0	71.75	89.60	88.43	88.50	89.60	86.98	89.60	86.98	71.75	70.25	70.25	70.73	70.25	70.33	70.25	70.33
Negatively Paired Means and Variances (Small N)																
0.0	28.38	30.78	28.45	29.18	28.03	28.25	31.40	28.23	28.38	26.58	28.20	28.03	28.20	28.05	27.23	28.10
0.2	30.20	33.83	30.90	31.80	31.10	31.58	34.43	31.50	30.20	28.98	29.98	29.80	30.03	29.90	29.23	29.90
0.4	31.28	37.98	32.35	34.20	32.95	33.75	38.28	33.75	31.28	30.05	31.08	30.83	31.00	31.00	30.38	31.00
0.6	33.80	46.88	37.68	41.15	38.90	39.60	46.48	39.98	33.80	32.00	33.38	32.78	33.13	33.10	32.08	33.08
1.0	34.20	51.48	43.68	47.60	48.35	44.40	51.60	46.30	34.20	31.68	32.98	32.58	32.53	32.70	31.73	32.50
Negatively Paired Means and Variances (Large N)																
0.0	70.58	70.60	70.70	70.78	70.65	70.60	71.28	70.60	70.58	70.50	70.55	70.53	70.55	70.58	70.50	70.58
0.2	69.78	72.23	70.48	70.88	70.78	70.60	72.50	70.60	69.78	69.73	69.78	69.78	69.78	69.78	69.73	69.78
0.4	68.58	75.10	71.80	72.93	73.08	72.05	75.05	72.35	68.58	68.25	68.45	68.35	68.38	68.43	68.25	68.43
0.6	68.58	81.08	80.13	80.10	80.58	78.90	81.08	78.90	68.58	67.88	67.95	67.95	67.88	67.95	67.88	67.95
1.0	66.43	87.38	87.20	86.98	87.38	85.78	87.38	85.83	66.43	65.08	65.08	65.13	65.08	65.15	65.08	65.15

[a] Each value in the table was based on 4,000 simulations.

Type I Errors of the Transformation Procedures

There were no conditions under which the use of the transformation procedures significantly improved upon the control of Type I error rates exhibited by the ANOVA F-test on untransformed data. These results are presented in Tables 17 and 18 for small and large sample sizes, respectively. As was discussed previously, no transformation can achieve homoscedasticity unless a functional relationship exists between the mean and variance of a distribution. Therefore, when no mean differences exist and there is heterogeneity of variance between treatment populations, a transformation cannot achieve homoscedasticity. Consequently the results obtained are not surprising.

Using a variance test to select between transformations (T_4) gave a slight improvement in control of Type I error rates when a liberal bias was present: however this improvement was not sufficient to warrant recommendation of the procedure. Also, when a conservative bias was present, use of this procedure served to accentuate the bias and therefore also led to reduced power.

As Type I error rates were not substantially improved by any of the transformation procedures, their power will not be discussed.

Table 17.

Type I Error Rates (%) for the ANOVA F-Test on Transformed and Untransformed Data

(Small N)

Group Size Condition	Coefficient of Variation	Normal Distribution					Chi-Square Distribution				
		F	T_1	T_2	T_3	T_4	F	T_1	T_2	T_3	T_4
u = 1.0	0.0	5.25	5.15	5.15	4.85	4.55	3.80	4.00	3.85	4.10	4.05
	0.2	5.15	4.85	4.95	5.10	4.90	4.50	4.55	4.70	5.10	4.80
	0.4	6.15	5.80	6.10	5.95	5.60	5.20	5.65	5.40	5.70	5.45
	0.6	5.85	5.90	5.65	6.10	5.45	6.25	6.95	6.75	7.85	6.40
	1.0	8.50	8.70	8.80	9.40	7.15	9.55	10.60	10.10	11.75	9.05
Unequal Group Sizes Positively Paired with Variances											
u = 1.5	0.0	4.10	4.00	4.00	4.10	3.95	4.35	4.40	4.60	4.90	4.70
	0.2	5.20	5.30	5.25	5.35	5.00	4.65	4.95	4.70	4.85	4.75
	0.4	4.30	4.20	4.25	4.20	3.80	4.10	4.60	4.20	4.90	4.50
	0.6	5.45	5.55	5.45	5.55	4.75	5.80	6.15	6.15	6.95	5.85
	1.0	5.85	5.55	5.55	6.15	4.95	7.80	9.35	8.65	11.50	7.55
u = 2.0	0.0	4.45	4.60	4.50	4.60	4.25	4.30	4.25	4.20	4.45	4.35
	0.2	4.05	3.95	3.95	3.85	3.70	4.10	4.15	4.05	4.20	4.25
	0.4	3.15	3.30	3.20	3.15	2.85	4.45	4.55	4.60	4.85	4.55
	0.6	3.40	3.40	3.40	3.20	3.00	3.60	4.00	3.75	4.95	3.95
	1.0	3.95	3.75	3.75	3.90	3.00	5.70	7.00	6.20	8.20	5.45
u = 3.0	0.0	5.00	4.90	4.95	5.15	4.60	4.55	4.45	4.55	4.45	4.50
	0.2	4.20	4.10	4.10	4.10	3.95	3.95	4.15	4.20	4.25	4.00
	0.4	2.80	2.55	2.85	2.65	2.55	3.35	3.55	3.40	3.60	3.50
	0.6	2.85	2.75	2.90	2.75	2.40	3.65	4.25	3.90	4.55	4.15
	1.0	3.25	3.45	3.20	3.45	2.50	4.25	5.10	4.60	6.55	4.25

Table 17. (continued)

Type I Error Rates (%) for the ANOVA F-Test on Transformed and Untransformed Data

(Small N)

Group Size Condition	Coefficient of Variation	Normal Distribution					Chi-Square Distribution				
		F	T_1	T_2	T_3	T_4	F	T_1	T_2	T_3	T_4
Unequal Group Sizes Negatively Paired with Variances											
u = 1.5	0.0	4.10	4.00	4.00	4.10	3.95	4.35	4.40	4.60	4.90	4.70
	0.2	5.15	5.30	5.20	5.35	5.00	4.75	4.75	4.80	5.00	4.85
	0.4	6.50	7.00	6.75	6.95	6.20	6.20	6.50	6.50	6.80	6.25
	0.6	8.85	9.55	9.25	9.90	8.25	7.05	7.85	7.30	8.75	7.35
	1.0	11.35	11.60	11.45	12.95	10.50	12.80	14.10	13.20	15.85	12.10
u = 2.0	0.0	4.45	4.60	4.50	4.60	4.25	4.30	4.25	4.20	4.45	4.35
	0.2	6.55	6.55	6.40	6.05	6.10	4.70	5.05	4.85	5.40	5.30
	0.4	8.75	8.90	8.80	9.00	8.05	8.45	8.35	8.20	9.05	8.50
	0.6	10.25	10.20	10.20	10.15	9.25	11.00	12.05	11.30	12.85	11.30
	1.0	15.85	16.05	15.65	17.35	13.85	16.00	17.80	16.95	20.20	15.35
u = 3.0	0.0	5.00	4.90	4.95	5.15	4.60	4.55	4.45	4.55	4.45	4.50
	0.2	5.75	6.20	5.80	6.15	5.45	5.50	5.95	5.65	6.20	6.05
	0.4	9.60	10.30	9.95	10.45	8.75	10.05	10.40	10.15	11.20	10.35
	0.6	13.55	13.70	13.45	14.25	12.60	13.65	15.50	14.65	15.75	13.20
	1.0	20.70	21.00	20.80	22.50	18.45	22.40	23.00	22.70	24.50	21.00

Table 18.

Type I Error Rates (%) for the ANOVA F-Test on Transformed and Untransformed Data

(Large N)

Group Size Condition	Coefficient of Variation	Normal Distribution					Chi-Square Distribution				
		F	T_1	T_2	T_3	T_4	F	T_1	T_2	T_3	T_4
	0.0	4.80	4.70	4.70	4.55	4.45	5.00	5.00	4.95	5.00	4.95
	0.2	5.90	6.25	6.10	6.70	5.45	4.90	5.20	5.15	5.85	5.50
u = 1.0	0.4	5.80	6.05	5.80	6.40	4.95	6.25	6.30	6.25	6.90	6.25
	0.6	6.35	6.25	6.40	7.75	6.10	5.30	6.30	5.95	8.45	5.65
	1.0	7.75	9.30	7.80	15.70	9.05	7.95	11.25	9.45	17.25	7.80
Unequal Group Sizes Positively Paired with Variances											
	0.0	4.95	4.75	4.65	4.55	4.30	4.00	4.00	3.90	4.35	4.15
	0.2	4.55	4.90	4.70	5.40	4.25	4.55	4.55	4.50	4.80	4.80
u = 1.5	0.4	5.60	5.35	5.30	5.55	4.20	5.15	5.35	5.20	6.05	4.95
	0.6	5.50	5.75	5.50	6.65	5.40	5.15	5.70	5.55	6.95	5.10
	1.0	5.75	7.60	6.60	12.05	6.40	6.10	8.95	7.20	14.70	6.10
	0.0	4.50	4.75	4.60	5.00	4.30	4.80	4.70	4.75	4.80	4.75
	0.2	3.55	3.50	3.45	3.70	3.05	4.00	4.50	4.35	4.60	4.40
u = 2.0	0.4	4.30	4.50	4.45	5.00	3.70	3.45	3.90	3.95	4.75	3.70
	0.6	2.65	2.75	2.65	4.00	2.55	3.35	4.05	3.55	5.35	3.30
	1.0	3.10	5.40	4.05	10.10	4.85	4.80	7.40	5.50	12.40	4.75
	0.0	4.75	4.75	4.80	4.65	4.40	5.30	5.30	5.25	5.55	5.45
	0.2	3.90	3.95	3.90	3.95	3.40	4.00	4.30	4.15	5.15	4.80
u = 3.0	0.4	3.10	3.15	3.20	3.85	2.70	2.50	2.90	2.65	3.60	2.45
	0.6	2.90	3.10	3.10	3.80	2.35	2.65	3.00	2.85	4.10	2.65
	1.0	3.10	3.90	3.15	7.20	3.10	3.40	5.80	4.40	10.95	3.40

Table 18. (continued)

Type I Error Rates (%) for the ANOVA F-Test on Transformed and Untransformed Data

(Large N)

Group Size Condition	Coefficient of Variation	Normal Distribution					Chi-Square Distribution				
		F	T_1	T_2	T_3	T_4	F	T_1	T_2	T_3	T_4
Unequal Group Sizes Negatively Paired with Variances											
u = 1.5	0.0	4.95	4.75	4.65	4.55	4.30	4.00	4.00	3.90	4.35	4.15
	0.2	6.00	6.40	6.25	6.55	5.25	5.60	5.45	5.40	5.85	5.75
	0.4	7.25	7.65	7.25	8.50	6.70	6.65	7.35	7.10	8.35	6.25
	0.6	8.45	10.05	9.00	12.70	9.35	7.80	8.35	8.15	10.30	7.80
	1.0	10.35	13.40	11.85	19.30	11.35	12.30	16.60	13.40	22.95	12.10
u = 2.0	0.0	4.50	4.75	4.60	5.00	4.30	4.80	4.70	4.75	4.80	4.75
	0.2	7.00	7.55	7.30	7.60	6.75	6.75	6.75	6.75	7.00	6.60
	0.4	6.80	7.30	6.75	8.20	6.05	7.85	8.90	8.20	10.15	8.35
	0.6	10.50	10.40	10.25	12.60	9.80	10.40	12.05	11.15	14.90	10.40
	1.0	14.95	16.45	15.70	20.80	13.00	16.25	19.00	16.80	24.45	15.75
u = 3.0	0.0	4.75	4.75	4.80	4.65	4.40	5.30	5.30	5.25	5.55	5.45
	0.2	8.50	8.50	8.50	8.40	7.25	5.70	5.65	5.75	5.95	6.00
	0.4	10.05	10.40	10.00	11.55	8.90	9.95	10.20	10.05	11.00	9.95
	0.6	13.20	14.30	13.45	16.95	12.20	14.90	16.25	15.70	19.45	14.80
	1.0	20.30	21.85	20.75	26.85	17.80	20.95	23.00	21.45	28.25	20.55

Concluding Remarks

Empirical Type I error rates for the tests of mean equality agree with the values obtained by other investigators (see Brown and Forsythe, 1974a and Kohr and Games, 1974). However, no effect of mean variance pairing on the relative power values of the Welch and F* tests was found in this study when the data was normally distributed, whereas both Brown and Forsythe (1974a) and Kohr and Games (1974) did observe such an effect. The discrepancy between these previous results and the results presented here may be explained by the absence of any particularly deviant means in the present study. Brown and Forsythe (1974a) showed that the Welch procedure had greater power than F* when extreme means had larger variances and vice versa for extreme means with smaller variances: they explained this by pointing out that means were weighted by n_j/s_j^2 in the Welch formula and by n_j in their formula for F*. Since there were no extreme means in the present study a large weighting factor at one end of the range of means would be exactly offset by a small one at the other end of the means range, especially so since the variances were also equally spaced over their range of values (except for c = 1.0).

Sampling from the chi-square population leads to bias in the sample means and variances as estimators of the corresponding population parameters and this is especially true when group sizes are small. In a positively skewed distribution the sample mean will tend to underestimate the population mean and this effect will be more pronounced

the greater the population variance and the smaller the sample size. These facts provide an explanation for why mean-variance pairing affected the power of all tests when the chi-square population was sampled. If larger means are associated with larger variances they will be underestimated to a greater degree than smaller means thus restricting the range of the means and leading to lower power for positive mean variance pairing: conversely when smaller means have larger variances they will be relatively more underestimated than larger means thus expanding the range of the means and leading to higher power for negative mean-variance pairing.

The effect of group size-mean pairing on W and F* is not easily explained because it is not easy to determine the relative degree of bias in the sample estimates that enter into the more complex formulae for these two statistics. However this effect diminishes at the larger sample size as does the effect of mean-variance pairing as would be expected if bias in the sample means is the reason for the power differences.

Type I error rates and power of F* were the least variable of the three tests. Thus F* is recommended if no information regarding population shape and variances is available. Type I error rates for F* only verged on the unacceptable when variance heterogeneity was at its highest value but they were still approximately equal to or less than those for the other tests. At the small sample size, although the power of F* was generally less than that of W, it was usually close to the a priori calculated power for F, and, on those occasions when it was less than this value, W did not perform substantially better. At the

larger sample size the power of F* again was usually below that of W but the only occasions on which it was much less than the a priori power of F, were those when variance heterogeneity was greatest (c = 1.0).

For the sequential procedures Type I error control was evaluated in terms of deviations from nominal alpha in both conservative and liberal directions. Conservative values were considered less acceptable because of the usual concomitant power loss for these conditions. The only situation in which a sequential procedure controlled Type I error rates better than any single means test was in the chi-square population when group sizes and variances were positively paired. However the power of the sequential procedures was not uniformly preferable to that of F*, and, since F* only had liberal Type I error rates at the highest degree of variance heterogeneity the advantage of performing the more complicated sequential procedure is negligible.

As had been predicted a priori, using data transformations to control Type I error rates in the presence of unequal group sizes and variances was completely ineffective. Thus Brown and Forsythe's (1974a) F* emerges as the only procedure that reliably controls Type I error rates despite non-normality, small sample size and heterogeneous variances and group sizes.

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., & Tukey, J.N. Robust Estimates of Location. Princeton, N.J.: Princeton University Press, 1972.

Aspin, A.A. An examination and further development of a formula arising in the problem of comparing two mean values. Biometrika, 1948, 35, 88-96.

Bartlett, M.S. Properties of sufficiency and statistical tests. Proceedings of the Royal Society, Series A, 1937, 160, 268-282.

Bartlett, M.S. The use of transformations. Biometrics, 1947, 3, 39-52.

Bartlett, M.S., & Kendall, D.G. The statistical analysis of variance heterogeneity and the logarithmic transformation. Journal of the Royal Statistical Society, Supplement 7, 1946, 128-138.

Behrens, W.U., Ein beitrag zur fehlerberechung bei wenigen beobachtungen. Landweise Jahresbuch, 1929, 68, 807-837.

Boneau, C.A. The effects of violations of assumptions underlying the t test. Psychological Bulletin, 1960, 57, 49-64.

Box, G.E.P. Non-normality and tests on variances. Biometrika, 1953, 40, 318-335.

Box, G.E.P. 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, 1954, 25, 290-302.

Box, G.E.P., & Andersen, S.L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society, Series B, 1955, 17, 1-26.

Box, G.E.P., & Cox, D.R. An analysis of transformations. Journal of the Royal Statistical Society, Series B, 1964, 26, 211-252.

Brown, M.B., & Forsythe, A.B. The small sample behaviour of some statistics which test the equality of several means. Technometrics, 1974a, 16, 129-132.

Brown, M.B., & Forsythe, A.B. Robust tests for the equality of variances. Journal of the American Statistical Association, 1974b, 69, 364-367.

Cochran, W.G. The distribution of the largest of a set of estimated variances as a fraction of their total. Annals of Eugenics, 1941, 11, 47-52.

Cohen, J. Statistical Power Analysis for the Behavioural Sciences. New York: Academic Press, Inc., 1969.

Curtiss, J.H. On transformations used in the analysis of variance. Annals of Mathematical Statistics, 1943, 14, 197-122.

Donaldson, T.S. Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. Journal of the American Statistical Association, 1968, 63, 660-676.

Draper, N.R., & Hunter, W.G. Transformations: Some examples revisited. Technometrics, 1969, 11, 23-40.

Fisher, R.A. The Fiducial argument in statistical inference. Annals of Eugenics, 1935, 5, 391-398.

Games, P.A., & Lucas, P.A. Power of the analysis of variance of independent groups on non-normal and normally transformed data. Educational and Psychological Measurement, 1966, 26, 311-327.

Games, P.A., Winkler, H.B., & Probert, D.A. Robust tests for homogeneity of variance. Educational and Psychological Measurement, 1972, 32, 887-909.

Gartside, P.S. A study of methods for comparing several variances. Journal of the American Statistical Association, 1972, 67, 342-346.

Glass, G.V., Peckham, P.D., & Sanders, J.R. Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 1972, 42, 239-288.

Hartley, H.O. The maximum F-ratio as a short-cut test for heterogeneity of variance. Biometrika, 1950, 37, 308-312.

Havlicek, L.L., & Petersen, N.L. Robustness of the t-test: A guide for researchers on effect of violations of assumptions. Psychological Reports, 1974, 34, 1095-1114.

Horsnell, G. The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika, 1953, 40, 128-136.

James, G.S. The comparison of several groups of observations when the ratio of population variances are unknown. Biometrika, 1951, 38, 324-329.

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

Layard, M.W.J. Robust large sample tests for homogeneity of variances. Journal of the American Statistical Association, 1974, 68, 195-198.

Levene, H. Robust tests for equality of variances. In I. Olkin et al. (Eds.) Contributions to probability and statistics, Stanford Press, 1960.

Levy, K.J. An empirical comparison of the Z variance and Box-Scheffe tests for homogeneity of variance. Psychometrika, 1975, 40, 519-524.

Linquist, E.F. Design and Analysis of Experiments in Education and Psychology, Boston: Houghton Mifflin, 1953.

Lunney, G.H. Using analysis of variance with a dichotomous dependent variable: An empirical study. Journal of Educational Measurement, 1970, 7, 263-269.

Marsaglia, G., MacLaren, M.D., & Bray, T.A. A fast procedure for generating normal random variables. Communications of the ACM, 1964, 7, 4-10.

Martin, C.G., & Games, P.A. ANOVA tests of homogeneity of variance when n's are unequal. Paper presented at the Annual Meeting of the American Educational Research Association, San Fransisco, April, 1976.

Miller, R.G. Jacknifing variances. Annals of Mathematical Statistics, 1968, 39, 567-582.

Mueller, C.G. Numerical transformation in the analysis of experimental data. Psychological Bulletin, 1949, 46, 198-223.

Olds, E.G., Mattson, T.B., & Odeh, R.E. Notes on the use of transformations in the analysis of variance. WADC Technical Report, 56-308, Wright-Patterson AFB, Ohio, 1956.

Overall, J.E., & Woodward, J.A. A simple test for heterogeneity of variance in complex factorial designs. Psychometrika, 1974, 39, 311-318.

Satterthwaite, F.E. Synthesis of variance. Psychometrika, 1941, 6, 309-316.

Satterthwaite, F.E. Synthesis of variance. Psychometrika, 1941, 6, 309-316.

Scheffe, H. The Analysis of Variance. New York: Wiley, 1959.

Scheffe, H. Practical solutions of the Behrens-Fisher problem. Journal of the American Statistical Association, 1970, 65, 1501-1508.

Schlesselman, J.J. Data transformation in the two-way analysis of variance. Journal of the American Statistical Association, 1973, 68, 369-378.

Spjøtvoll, E., & Stoline, M.R. An extension of the T-method of multiple comparison to include the cases with unequal sample sizes. Journal of the American Statistical Association, 1973, 68, 975-978.

Tarter, M.E., & Kowalski, C.J. A new test for the class of transformations to normality. Technometrics, 1972, 14, 735-744.

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

Welch, B.L. The significance of the difference between two means when the population variances are unequal. Biometrika, 1938, 29, 350-362.

Welch, B.L. The generalization of Student's problem when several different population variances are involved. Biometrika, 1947, 34, 28-35.

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

APPENDIX

Glossary of Statistical Terms Not Defined in Text

DEGREES OF FREEDOM: A sample of n variate values, X_i, is said to have n degrees of freedom, whether the variates are dependent or not, and a statistic calculated from it is, by natural extension said to have n degrees of freedom. But, if K functions of the sample values are held constant, the number of degrees of freedom is reduced by K. For example, the statistic $\sum_{i=1}^{n} (X_i - \bar{X})^2$, where $\bar{X}$ is the sample mean, is said to have n-1 degrees of freedom, since the sample mean is regarded as fixed.

By a further extension, the distribution of a statistic based on n independent variates is said to have n degrees of freedom, particularly in relation to $X^2 = \sum_i Z_i^2$ (see p. 50)

KURTOSIS: A term used to describe the extent to which a unimodal frequency curve is 'peaked'; that is to say, the extent of the relative steepness of ascent in the neighbourhood of the mode. The moment ratio

$$\beta_2 = E(X_i - \mu)^4 / (E(X_i - \mu)^2)^4$$

is used as a measure of kurtosis and is related to γ_2, the measure used here, by $\gamma_2 = \beta_2 - 3$. If γ_2 is adopted as a measure of kurtosis, the value it assumes for a normal distribution, namely zero, is taken as a standard. Curves for which the ratio is less than, equal to or greater than zero are known respectively as platykurtic, mesokurtic and leptokurtic. Thus a platykurtic distribution is flatter or less peaked than a normal distribution, whereas a leptokurtic distribution is more so.

MOMENT: A moment is the expected value (mean value) of the power of a variate. For example, $E(X_i - \mu)^4$ is the expected value of the fourth power of deviations from the mean and is known as the fourth moment about the mean.

NULL HYPOTHESIS: In general this term relates to a particular hypothesis under test, as distinct from the alternative hypotheses which are under consideration. It is therefore the hypothesis which determines the probability of the Type I error. Here the term is restricted to a hypothesis under test of 'no difference'. Thus the null hypothesis in a test of mean equality is that 'no mean differences are present'.

SAMPLING DISTRIBUTION: The distribution of a statistic in all possible samples which can be chosen according to a specified sampling scheme. The expression always relates to a sampling scheme involving random selection, and most usually concerns the distribution of a function of a fixed number n of independent variates.

SKEWNESS: A term for assymmetry in relation to a frequency distribution. If a unimodal distribution has a longer tail extending towards lower values of the variate it is said to have negative skewness; in the contrary case, positive skewness.