ED 255 571	TM 850 214
AUTHOR	Olejnik, Stephen F.; Algina, James
TITLE	Power Analysis of Selected Parametric and Nonparametric Tests for Heterogeneous Variances in Non-Normal Distributions.
PUB DATE	2 Apr 85
NOTE	42p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Computer Simulation; *Hypothesis Testing; Sampling; *Statistical Distributions; Statistical Studies; Transformations (Mathematics)
IDENTIFIERS	*Power (Statistics); Type I Errors; *Variance (Statistical)

ABSTRACT

The present investigation developed power curves for two parametric and two nonparametric procedures for testing the equality of population variances. Both normal and non-normal distributions were considered for the two group design with equal and unequal sample frequencies. The results indicated that when population distributions differed only in scale, the Klotz procedure consistently provided the most sensitive test of variance equality. With symmetric distributions which differ in both mean and scale the Klotz procedure with mean aligned data was the most powerful procedure. When sampled distributions were skewed and differed in mean as well as scale the O'Brien or Brown-Forsythe procedures were preferred. (Author)

Power Analysis of Selected Parametric and Nonparametric Tests for Heterogeneous Variances in Non-Normal Distributions

Stephen F. Olejnik

James Algina

University of Florida

Paper presented at the American Educational Research Association meetings April 2, 1985, Chicago, Illinois.

Precis

The present investigation developed power curves for two parametric and two nonparametric procedures for testing the equality of population variances. Both normal and non-normal distributions were considered for the two group design with equal and unequal sample frequencies. The results indicated that when population distributions differed only in scale, the Klotz procedure consistently provided the most sensitive test of variance equality. With symmetric distributions which differ in both mean and scale the Klotz procedure with mean aligned data was the most powerful procedure. When sampled distributions were skewed and differed in mean as well as scale the O'Brien or Brown-Forsythe procedures were preferred.

Power Analysis of Selected Parametric and Nonparametric Tests for Heterogeneous Variances in Non-Normal Distributions

Until recently tests of scale (variance) have been viewed by many social scientists as being of minor interest. This disinterest may be partially attributed to introductory and intermediate statistics textbooks which discuss these procedures briefly and only in the context of determining whether the homogeneity of variance assumption of the independent samples t-test and ANOVA F-test is met. The importance of this preliminary test of scale is often further minimized by citing analytic or empirical investigations that have indicated that when sample sizes are equal, parametric tests of means are robust to heteroscedasticity (Glass, Peckham and Sanders, 1972). The robustness of parametric tests of location however is not without controversy. Based on the exact Type I error rates, Ramsey (1980) challenged the robustness claim and provided guidelines to conditions under which the t-test is robust to heteroscedasticity. When sample sizes differ, considerable evidence supports the conclusion that parametric tests of location can be either liberal or conservative depending on the relationship between sample sizes and variances. Under these conditions tests of scale are generally viewed to be particularly important. New interest in tests of variance equality has also recently developed as a result of new theories in the social sciences. Games, Winkler and Probert (1972) cited several areas of study in which the comparative analysis of distribution dispersion are of primary interest. A program intervention may, for example, have the effect of reducing group variability while having little or no effect on the population mean. A new development in this area has been the introduction of statistical procedures designed to investigate the simultaneous influence of multiple independent variables on variance equality in factorial designs (Games, 1978).

In statistics texts, the most frequently cited analysis procedures for testing the equality of group variances are those suggested by Bartlett (1937), Cochran (1941) and Hartley (1950). These strategies however have been shown analytically and empirically to be extremely sensitive to the shape of the population distribution (Box, 1953; Games et al., 1972; Layard, 1973; Overall and Woodward, 1974). In particular it is the kurtosis of the distribution which affects the standard error of the variance and the distribution of the Bartlett, Cochran, and Hartley test statistics. When the population is leptokurtic a liberal test of the variance will occur whereas with a platykurtic distribution the test will be conservative. One solution to the problems caused by non-normality is to adjust the test statistic to take into account the effects of non-normality. Box and Andersen (1955) using permutation theory developed a procedure which adjusts Bartlett's M using a sample estimate of the population kurtosis. Studies of this approach have had mixed results. Games et al. (1972) using empirical methods found Box and Andersen's statistic to provide liberal hypothesis tests ($\sim$ .08) when the distribution was normal or skewed and to provide conservative hypothesis tests when the distribution was uniform. With a leptokurtic distribution however the actual Type I error rate was similar to the nominal Type I error rate. Miller (1968) on the other hand found the Box-Andersen statistic to be robust to normal, light tailed and heavy tailed distributions. The differences in results may be attributed to sample sizes; Games et al. considered samples of 18 per group while Miller studied samples of 25 per group. The relatively large sample size needed and the computational difficulty have been viewed as limitations for the Box-Andersen approach. In addition the procedure cannot be used for complex factorial designs.

An alternative to adjusting the test statistic is to construct a dependent variable that measures variability and to conduct an ANOVA on the constructed dependent variable. For example, Box (1953) building on the work of Bartlett and Kendall (1946) suggested dividing the sample into subsamples and computing separate estimates of variance for each subsample. To test the equality of group variances the subsample variances are transformed using a log transformation and log S^2 is used as the dependent measure in computing the ANOVA F-ratio. This approach has the advantage of being applicable to complex factorial designs. Provided that the parent distribution for the cells of the design have the same kurtosis, that the number of observations is equal for all cells, and that the same number of observations comprise each subsample, then the homoscedasticity assumption will be approximately met. While the distribution of log S^2 can be non-normal, under the sample size restrictions listed a'ove ANOVA is reasonably insensitive to non-normality. Several investigations have considered the procedure in situations involving various population distributions. The results of these studies have shown that the log transformation is robust to the normality assumption (Box, 1953; Games et al., 1972; Layard, 1973; Overall and Woodward, 1974; Levy, 1975; Martin and Games, 1977; Ga..es, Keselman and Clinch, 1979). These same investigations have also shown that Box's procedure is less powerful than other appropriate procedures when the distribution is normal, uniform, skewed or leptokurtic. In addition to lacking statistical power the log transformation has two additional limitations. Fitst, the procedure does not specify the number of subgroups which should be formed. Martin and Games (1975) however have found that the whole number closest to the equare root of the group sample size to provide optimal power. A second more serious limitation is that results of the analysis can vary depending on the formation of the subgroups. Two researchers using the same data could arrive at opposite conclusions. This lack of uniqueness of results has led several researchers to abandon the approach (Brown and Forsythe, 1974; O'Brien, 1978).

Levene (1960) suggested several transformations of the original data in order to test the equality of group variance. One approach was to calculate the square of the difference between each observation and its group mean [$p_{ij} = (x_{ij} - \bar{x}._j)^2$]. Substituting the p_{ij}'s for the original data an ANOVA F-ratio is computed. Empirical studies of this approach have been mixed and inconsistent (Miller, 1968; Games et al., 1972). Miller's results suggest that with relatively large sample sizes (n=25) the approach is robust to normal, light tailed and heavy tailed distributions. Games et al. results suggest that with small sample sizes (n=6) the test is liberal with normal, skewed and uniform distributions. These results are consistent with a theoretical analysis reported by O'Brien (1978). The approach was also shown to have slightly lower power than the Box-Andersen statistic but greater power than the log transform approach (Miller, 1968).

A second transformation suggested by Levene (1960) was to calculate the absolute value of the difference between each observation and its group mean ($Z_{ij} = |x_{ij} - \bar{x}._j|$). The absolute differences replace the original data and the ANOVA F-ratio is computed. Miller (1968) discarded this approach arguing that the test statistic is not asymptotically distribution-free. Others have considered this approach however and found mixed results. Games et al. (1972) found the approach to provide a liberal hypothesis test for normal, skewed, uniform and leptokurtic distributions. Brown and Forsythe (1974) using larger sample sizes provided evidence suggesting that the approach was appropriate for normal and leptokurtic distributions but liberal for distributions which were both leptokurtic and skewed. Their results also indicated that Levene's absolute difference transformation had statistical power similar to the F-ratio when the distributions were normal. With leptokurtic distributions some power was lost but the procedure was more powerful than other robust competing strategies.

A simple test of variance equality, suggested by Overall and Woodward (1974), involves transforming the group variance to a Z statistic. The Z variance test can be used in complex factorial designs but studies have indicated that it is not robust to non-normality (Levy, 1975). With normal distribution however the procedure provides appropriate Type I error rates and has slightly greater power than Box's log transform approach (Overall and Woodward, 1974; Levy, 1975; Martin, 1976).

Miller (1968) suggested an application of the jackknife technique to testing the equality of variances in two groups. A generalization of this approach to multiple groups was later presented by Layard (1973). In one application of this approach, the group variance (S_j^2) as well as the variance of each of the n subgroups (S_{ij}^2) created by deleting one observation at a time from the group are computed. For each subgroup a new variable is created by subtracting (n-1) times, the log of the subgroup variance from n times the log of the group variance [$p_{ij} = n \log S_j^2 - (n-1) \log S_{ij}^2$]. The p_{ij} variable is then used as the dependent measure in calculating the ANOVA F-ratio. Unlike Box's log transformation, this jackknife provides a unique solution for a given data set. However, if large samples are studied more than one observation may be deleted in creating the subgroup and the unique solution is no longer available. The approach has been studied in both simple and multiple factor designs with the deletion of one observation and the deletion of several observations. The results have indicated that with heavy tailed distributions the test is liberal and with light tailed distributions the test is conservative (Miller, 1968; Layard, 1973; Brown and Forsythe, 1974; Martin and Games, 1977; O'Brien, 1978; Games et al., 1979). Although power analyses have been conducted their usefulness is doubtful given the lack of robustness of the procedure.

Brown and Forsythe (1974) in an approach similar to Levene's absolute difference from group means suggested transforming the original data by using the absolute difference of each observation from its group median: $\lambda_{ij} = \left| X_{ij} - M_j \right|$). The transformed variable is then used in the calculation of the ANOVA F-ratio. The procedure suggested by Brown and Forsythe is attractive because of its computational simplicity and its versatility. The absolute differences from the group median can easily be used in complex factorial designs. In studying the Type I error rate of this procedure researchers have found that it provides a conservative test of variance with normal and light tailed distributions (O'Brien, 1978; Games et al., 1979). With larger sample sizes appropriate Type I error rates were observed by Brown and Forsythe (1974). With heavy tailed distributions appropriate Type I error rates were obtained for both small and large sample sizes(Brown and Forsythe, 1974; O'Brien, 1978; Games et al., 1979). With regard to statistical power the Brown-Forsythe test was less powerful than other competing strategies such as the F-ratio, both of Levene's tests, and the jackknife technique when the distribution was normal. It was however more powerful than the log transformation. With heavy tailed distributions the Brown-Forsythe statistic was more powerful than the log transform but less powerful than Levene's absolute difference from the group mean.

Recently, O'Brien (1978) suggested a new strategy for testing the equality of group variances. His solution involves transforming the original scores by taking into consideration both the squared deviation of the score from the group mean and the group variance. Specifically the transformation is:

$$r_{ij} = [\ (\omega + n_j - 2)\ n_j\ (y_{ij} - \bar{y}._j)^2 - \omega s_j^2\ (n_j-1)]\ /\ [(n_j-1)(n_j-2)]$$

where n_j is the number of observations in the j^{th} group, y_{ij} is the i^{th} observation in the j^{th} group, $\bar{y}._j$ is the average score of the j^{th} group, and ω is a weighting factor. When ω is set equal to 0 the transformed variable is a modification of Levene's squared difference from the group mean transformation $r_{ij}(0) = [\frac{n_j}{n_j-1}\ (y_{ij} - y._j)^2\ ]$ and when ω=1, O'Brien's statistic is similar to Miller's jackknife statistic $[r_{ij}(1)=n_j s_j^2-(n_j-1)s_{ij}^2]$ (O'Brien, 1979). O'Brien (1981) recommends however for most situations that ω=.5 resulting in $r_{ij}(.5)=[(n_j-1.5)n_j(y_{ij}-\bar{y}._j)^2-.5S_j^2(n_j-1)]/(n_j-1)(n_j-2)]$. The r_{ij} variable is then used as the outcome measure in calculating the ANOVA F-ratio. The procedure has the advantage of being easily calculated and can be applied to complex factorial designs.

In an empirical investigation of the properties of this approach to testing variances, O'Brien generated data for a 4 x 3 factorial design. The results indicated that the r_{ij} (.5) provided a conservative hypothesis for normal and light tailed distributions. With a heavy tailed exponential distribution Type I error rates similar to the nominal level were observed. In comparing this approach with several alternatives including Levene's squared difference transformation, Brown and Forsythe's absolute differences from the group median and Box's log transformation, O'Brien concluded that the r transformed variable and the Brown-Forsythe statistic provide the best alternatives of those studied with Brown and Forsythe's approach being preferred with heavy tailed distributions and the r-transform preferred with normal and light tailed distributions.

The sensitivity to non-normality shown by many of the variance tests discussed above has led some researchers to develop nonparametric tests of scale. Two nonparametric tests of scale frequently cited in nonparametric textbooks (Lehmann, 1975; Marascuilo and McSweeney, 1977) were developed by Siegel and Tukey (1960) and Klotz (1962). The Siegel-Tukey rank test required ranking the pooled data from two samples by assigning a rank of 1 to the lowest observation, a rank of 2 and 3 to the highest and second highest observations, respectively. The ranking continues by alternating the assignment of ranks from the two ends of the distribution. The ranks are then analyzed using the Wilcoxon rank test. Exact tables are available for small sample problems and for large samples a Z test is recommended including a correction for ties. Alternatively the Kruskal-Wallis statistic can be applied to these ranks to generalize the procedure to situations involving more than two groups (Puri, 1964).[1] Klotz's (1962) test is a normal scores approach for comparing distributions in which the data from two samples are pooled and ranked from lowest to highest. The assigned ranks are then replaced by their inverse normal score $[\phi^{-1}(\frac{i}{N+1})]$. The test statistic is calculated using the squares of the inverse normal scores. A large sample form of the test, that can be used with two or more groups, involves calculating (N-1) SSB/SST (Puri, 1964).[2] Here SSB and SST denote sums of squares between and total, respectively. Klotz showed that his test is more efficient than the Siegel-Tukey test for normal and light tailed distributions while the Siegel-Tukey statistic is preferred for heavy tailed distributions.

Critics of the rank tests of scale have argued that this approach is of limited value since it may be sensitive to between group differences in the median as well as to between group differences in variance (Moses,1963). Miller (1968) for example rejects the approach on this basis. It has been suggested however that sample estimates of location could be used to align the data using the group mean or median before the ranking process begins (Lehmann, 1975; Marascuilo and McSweeny, 1977). Criticism of the ranking procedure has been based on asymptotic theory and small sample properties of these statistics have not been reported in the literature. The results of the effects of alignment using the sample mean or median have also not been reported in the literature.

The purpose of the present investigation was to develop and compare statistical power curves for several parametric and nonparametric tests of scale for normal, light tailed and heavy tailed population distributions. Previously, power studies of parametric tests of variance equality have not considered the nonparametric alternatives. In addition discussions of nonparametric tests of scale have been based on asymptotic behavior of these statistics and little has been published regarding the small sample properties of these procedures. In particular the effect of adjusting for differences in the location parameters between populations has not been considered. The procedures suggested by Brown and Forsythe (1974) and O'Brien (1978) have been selected to represent the parametric tests of variance equality. These procedures were chosen since previous studies have shown that they are: 1) relatively insensitive to distributional non-normality, 2) as powerful or more powerful than competing approaches, 3) can be used in factorial designs and 4) easy to compute and therefore attractive to applied social science researchers. The rank tests of scale considere were those developed by Siegel and Tukey (1960) and Klotz (1 These procedures were considered since they are 1) familiar to many data analvsts, 2) relatively more efficient than other competing nonparametric tests of scale, 3) applicable to a single factor design with multiple groups (Puri, 1960), and 4) easy to compute.

In considering these four tests of scale the following questions were of particular interest:

1) When populations have a common location parameter but differ in scale, which of the four procedures will provide the most sensitive test for that difference?

2) When populations differ with regard to their means, what effect do these differences have on the Type I error rates of the nonparametric tests of scale for small samples?

3) When population mean differences exist how do the nonparametric tests based on the aligned data compare to the parametric tests using the unaligned data?

Method

Although the procedures consiaered in the present paper are applicable to multiple group designs, it was decided to make the power comparisons based on the analysis of two groups. This restriction was made to conserve resources in order to consider multiple levels of other factors thought to have a greater effect on the power curves. In generating the power curves four parameters were manipulated: 1) sample size, 2) form of the parent distribution, 3) means of the parent distribution and 4) variance of the parent distribution.

Sample Size. Samples of (10,15); (15,10); (20,20); (17,23); and (23,17) were included in the investigation. The sample sizes were considered to be moderate and representative of those frequently found in research studies in the social sciences. The small departures from equal n were chosen specifically to reflect a small loss of subjects often found in social research.

Distribution Form. A normal and five non-normal parent distributions were considered. The non-normal distributions included a symmetric platykurtic (light tailed) distribution, a symmetric leptokurtic (heavy tailed) distribution, a slightly and a moderately skewed distribution, and a distribution which was both skewed and leptokurtic. The population characteristics of these distributions are discussed in the data generation section below.

Population Means. The simulations considered populations having a common mean as well as populations with means that differed by .2, .5 or .8 standard deviation units when variances were equal. When variances were unequal, differences in population means were equal to .2, .5 or .8 pooled standard deviation units. These effect sizes conform to what Cohen (1977) has suggested as guidelines defining small, medium or large effects.

Population Variances. To study the Type I error rates of the procedures under consideration data from populations with equal variances were generated. To study the sensitivity of the procedures to unequal variances data were generated from populations having the following variance pairs: (1,1.5); (1,2.0); (1,2.5); (1,3.0); (1,3.5); (1,4.0). The choice of these variance differences was based on two considerations. First it was believed that the conditions considered reflected actual situations encountered by applied researchers. And second it was believed that with unequal sample sizes differences in the variance of the magnitude considered here would affect the Type I error rate of the independent sample's t-test. To support this belief a brief simulation study was conducted in which data were generated from five distribution forms, five sample size combinations and seven levels of variance difference. Table 1 reports the observed Type I error rate for an independent sample t-test when the nominal significance level was .05.

Insert Table 1 about here

The results indicate that as expected when the sample sizes and variances are inversely related and the sample size is small a liberal test occurs with variance ratios as small as 1:1.5. With a direct relationship a conservative test occurs with ratios as small as 1:3.0. With larger sample sizes the problem is not as great. However when sample size and variance are inversely related an inflated Type I error rate was observed with variance ratios of 1:2.5 or smaller. With equal sample sizes, differences in group variances have no serious effect for the sample size considered.

Data Generation. Data for the study were generated using the SAS computing package. Scores on the dependent measure were created based on the linear model function $Y_{ij} = \mu..+\alpha._j + \sigma_j \varepsilon_{ij}$, where Y_{ij} is the i^{th} observation in the j^{th} group. The grand mean $\mu..$ was set equal to 10. The effect size parameter for the j^{th} group, $\alpha._j$, was 0, .2, .5 or .8 pooled standard deviation units to study the effect population mean difference. To generate the random error component the SAS NORMAL function was used to generate observations on a standard normal random variable, X_{ij}. To study normal distributions ε_{ij} was set equal to X_{ij}. To study the effect of non-normality X_{ij} was transformed using the power function suggested by Fleishman (1978): $\varepsilon_{ij} = [(dX_{ij} + c) X_{ij} + b] X_{ij} + a$. The constants a, b, c and d are chosen to transform the normally distributed variable to a variable with known skewness and kurtosis and mean zero and variance one. Five non-normal distributions were considered in the present study. The frequency distribution at half standard deviation invervals and descriptive statistics are reported in Table 2. Values reported in Table 2 are based on 20,000 observations generated for each distribution. The coefficient σ_1

Insert Table 2 about here

---

was set equal to one for all observations in group one. Thus the variance in group one was equal to one for all conditions. The coefficient σ_2 was chosen so that the variance of the second group was increased from 1 to 4 in increments of .5 units.

Computed Test Statistics. For each sample generated the statistics developed by O'Brien (1978), Brown and Forsythe (1974), Klotz (1962) and Siegel and Tukey (1960) were computed.

O'Brien's transformation with $\omega = .5$ was used. Each observation within each group was transformed using the following equation:

$$r_{ij} = \frac{(n_j-1.5)n_j \ (y_{ij}-\bar{y}_{.j})^2 - .5S_j^2 \ (n_j-1)}{(n_j-1) \ \ n_j-2)} .$$

With the transformed variable as he dependent measure the usual ANOVA F-ratio was completed:

$$F_{OB} = \frac{\sum_j n_j \ (\bar{r}_{.j} - \bar{r}_{..})^2 \ / \ (J-1)}{\sum_i \sum_j (r_{ij} - \bar{r}_{.j})^2 \ / \ (N-J)} .$$

The critical test statistic has J-1 and N-J degrees of freedom.

Brown and Forsythe's statistic was calculated after determining the absolute difference between each observation and the median observation of its group $\lambda_{ij} = \left| Y_{ij} - M_j \right|$. The computed test statistic was an ANOVA F-ratio with λ_{ij} as the dependent measure:

$$F_{BF} = \frac{\sum_j n_j \, (\bar{\lambda}_{.j} - \bar{\lambda}_{..})^2 \, / \, (J-1)}{\sum_i \sum_j (\lambda_{ij} - \bar{\lambda}_{.j})^2 \, / \, (N-J)} \quad .$$

The critical test statistic has J-1 and N-J degrees of freedom.

Siegel and Tukey's statistic was calculated after ranking the combined observations so that the lowest observation received a rank 1, the highest and second highest received a rank of 2 and 3 respectively and so forth until all observations were ranked. The test statistic using Kruskal and Wallis's formula for comparing mean ranks (R_j) was used:

$$H = \frac{12}{N(N+1)} \sum_j \frac{R_j^2}{n_j} - 3\,(N+1).$$

The H test statistic is asymptotically distributed as chi-square with J-1 degrees of freedom.

Klotz's procedure requires the ranking of the total sample across groups from 1 to N. The rank data are then replaced with normal scores $Z_{ij} = \phi^{-1} \left(\frac{1}{N+1}\right)$. The test statistic is then calculated as:

$$K = (N-1) \frac{\sum_j n_j \left(\bar{Y}_{.j} - \bar{Y}_{..} \right)^2}{\sum_i \sum_j Y_{ij}^2 - \frac{1}{N} \left(\sum_i \sum_j Y_{ij} \right)^2} \quad ,$$

where $Y_{ij} = Z_{ij}^2$. K has an asymptotically distributed chi-square with J-1 degrees of freedom.

For each condition studied, 1000 replications of the four statistics were computed and the frequency at which each procedure rejected the null hypothesis of equal variance at the .05 and .10 level were recorded.

Results

The results of the simulation are reported in two sections. The first section presents the results for the case in which the mean of the parent distribution did not vary across simulated treatment groups. This section is divided into two parts. Part one presents the Type I error rate of each procedure. The second part presents the power results. The second section presents the results for the case in which the simulated treatment groups had an expected mean difference of .2, .5 or .8 pooled standard deviation units. The first part of this section reports the Type I error rates and the second part compares the power curves for the four strategies. This section also includes an analysis of the effect of adjusting for sample differences in means and medians. To conserve space only the results at a nominal .05 level of significance are reported. Similar results were obtained at the nominal .10 significance level. In evaluating the robustness of each procedure, it was decided that observed proportions of Type I errors two standard errors above or below the nominal significance level would be judged as unacceptable. Based on 1000 replications the standard error for a nominal .05 significance level is .0069, so observations outside the interval (.036, .064)were considered either less than or greater than the nominal significance level.

Common Means

Type I error rates observed for the four procedures under consideration are reported in Table 3. With the exception of O'Brien's procedure, when used with a leptokurtic distribution, all of the procedures appeared to be insensitive to the form of the parent distribution. With the leptokurtic distribution, O'Brien's statistic consistently resulted in Type I error rates that were less than the nominal significance level. These results are consistent with those reported by O'Brien (1978).

- -

Insert Table 3 about here

- -

Power. The power curves obtained for the tests of variance equality with sample sizes of 20/20, 17/23 and 23/17 are reported in Table 4. The results for samples of 10/15 and 15/10 are not reported here since the relationships between the competing analysis strategies are similar to those presented for the sample sizes 17/23 and 23/17. With smaller samples however the proportion of hypotheses rejected are considerably lower. For example with sample sizes of (15,10) and a normal population O'Brien's statistic rejected 49.3 percent of the hypotheses when the variances differed by a ratio of 1 to 4.

- -

Insert Table 4 about here

- -

The results reported in Table 4 indicate that power curves based on samples of 20/20 and 23/17 to be very similar, whereas the power estimates based on samples of 17/23 were somewhat lower than for the other two sample size combinations. The ordering of the tests, in terms of power, however was very similar for all three sample size combinations. Table 5 exhibits summary partial orders, in terms of the power of the four procedures. These are somewhat idealized since the ordering is not precisely the same for every combination of ratio of variance and sample size. Nevertheless the partial orders are generally accurate as summaries of the results. The

- -

Insert Table 5 about here

- -

partial order for the normal distribution, for example, indicates that the O'Brien, Brown-Forsythe and Klotz tests are typically equivalent in terms of power and are superior to the Siegel-Tukey tesc. It appears that when populations differ only in their scale but are identical in their form and location parameter, then Klotz's approach consistently provides the most sensitive test of variance equality. Under specific distributions however one or more of the other strategies may provide comparable power.

Unequal Means

Type I errors. The rank tests of scale have been challenged as being inappropriate when populations differ with respect to their location parameter (Moses, 1963). Several authors have suggested solving this problem by aligning the data using an estimated group location parameter. Table 6 presents the actual Type I error rates for the mean and median aligned data when the populations differed by a small (.2), medium (.5) and large (.8) shift parameter. All results are for a nominal alpha level of .05. Only the results for samples of 20/20 and 23/17 are reported. Liberal tests are identified by a * and conservative tests by a t. For the most part similar results were obtained from the 17/23 sample size combination. The main exception occurred with the moderately skewed distribution. The unaligned Klotz test was quite liberal with this distribution.

- -

Insert Table 6 about here

- -

The effect of differences in population means on the actual Type I error rate was fairly similar for equal and unequal sample sizes. The results indicate that for symmetric distributions (normal, platykurtic and leptokurtic) the best control over Type I error rates is achieved using the mean aligned Siegel-Tukey or Klotz test. The median aligned Siegel-Tukey test has a tendency to be liberal with unequal sample sizes. The other tests have a tendency to be conservative, especially for larger effect sizes. For the slightly skewed distribution (skewness = .50, kurtosis = 0.0) the unaligned Siegel-Tukey and Klotz tests tend to have the best control over Type I error rates. However the mean aligned Klotz test also exhibited reasonable control of Type I errors for both equal and unequal cell frequencies, and the median aligned Siegel-Tukey test worked well with equal cell frequencies. For the moderately skewed (skewness = .75, kurtosis = 0.0) and for the skewed and leptokurtic distribution (skewness = 1.75, kurtosis = 3.75) none of the nonparametric procedures have adequate control over the Type I error rates.

<u>Power</u>. The estimated power curves for the original four tests of scale plus the four Siegel-Tukey and Klotz tests based on aligned data are reported in Table 7, 8, and 9 for samples of 20/20, 23/17 and 17/23 respectively. With symmetric distributions the unaligned Siegel-Tukey and Klotz procedures always became more conservative as the shift parameter increased. As a result the power to detect scale differences tended to decrease as the shift parameter increased and so the power curves for these procedures are reported only for those conditions where there was a small difference in population means. Aligning the data using the sample mean or median often provided an acceptable solution to the problem of population mean differences. As a result the power curves for the aligned Klotz and and Siegel-Tukey procedures were very similar across the three levels of the shift parameter considered. In addition the shift had no effect on the power

- -

Insert Tables 7, 8 and 9 about here

- -

curves for the Brown-Forsythe and O'Brien statistics. The power curves reported for those procedures which were unaffected by the shift are based on the average proportion of the hypotheses rejected across the small, medium and large shift parameter at each level of the variance ratio. Finally power results number are not reported for situations in which a Siegel-Tukey and/or Klotz test was liberal.

With symmetric distributions the mean adjusted Klotz test tends to be the most powerful procedure based on ranks. For these distributions, Table 10 exhibits partial orders of the mean adjusted Klotz test, the O'Brien test and the Brown-Forsythe test. With symmetric distributions the procedure of choice is the Klotz test; in all cases it is either more powerful than the Brown-Forsythe and O'Brien tests or has power equivalent to the O'Brien test.

- -

Insert Table 10 about here

- -

For slightly skewed distributions the procedures suggested by O'Brien, Brown and Forsythe, and the median aligned Klotz test provided similar power curves. The aligned Klotz procedure also provided comparable power estimates when the difference between population means was small. These results were consistent across equal and unequal sample size combinations.

For distributions which were both skewed and leptokurtic the unaligned and aligned rank tests of scale were quite liberal and therefore power results are not reported. A comparison of the power curves for O'Brien and Brown-Forsythe statistics indicated that when sample sizes were equal the Brown-Forsythe had a slight power advantage. When sample sizes were unequal there appeared almost no difference in the sensitivity of the tests.

The preceeding results apply to only a subset of the possible relationships among cell frequency, size of variance, and size of mean. For the unequal n case, there are four possible relationships to investigate: 1) Larger frequency, variance and mean in the same cell; 2) Smaller frequency, larger variance and mean in the same cell; 3) Larger frequency, variance and smaller mean in the same cell; and 4) Smaller frequency, larger variance and smaller mean in the same cell. The preceeding results for the 17/23 conditions are for relationship 1, whereas the results for the 23/17 are for relationship 2. To investigate whether the latter two relationships impact the power results, the simulations were repeated, but only for the symmetric distributions. The power order for situation 3 was quite similar to that for situation 1, whereas the ordering for situation 4 was quite similar to situation 2. Apparently the relationship between cell frequency and cell variance has a small impact on the power order of the Brown-Forsythe, mean aligned Klotz and O'Brien tests. However neither the relationship between cell variance and cell mean had an effect on the ordering. For the equal n case there are two possible solutions: 1) Larger variance and mean in the same cell; and 2) Larger variance and smaller mean in the same cell. The preceeding results suggested that the relationship between cell variance and mean did not impact the power order and that the impact of the cell frequency - cell variance relations was quite minor. Therefore additional simulations were not undertaken to investigate the impact of situation 2.

Summary and Conclusions

Based on the preceeding results, the following conclusions can be set forth:

1. When sampling from two populations that have identical shapes and means, the Brown-Forsythe, Klotz, O'Brien and Siegel-Tukey tests have actual Type I error rates near the nominal alpha level. For all the distributions investigated, the Klotz test had power equal to or greater than the power of the other tests. For the normal and platykurtic distributions, these results are consistent with asymptotic results indicating that the Klotz is more efficient than the Siegel-Tukey. However for the leptokurtic distribution the small sample results are not consistent with the large sample theory.

2. The results support O'Brien's conclusion that, with normal and light tailed distributions, his test is more powerful than the Brown-Forsythe test; with heavy tailed distributions it is less powerful. Because neither test is affected by differences in means these results obtain in the conditions with equal means and the conditions with unequal means.

3. As the differences between means increases, the unaligned Klotz and Siegel-Tukey tests become quite conservative and there is a concomitant reduction in power. Fligner (1979) investigated the Siegel-Tukey test and reported a similar trend. However, because Fligner studied smaller mean differences he did not demonstrate the excessively conservative tendency of the Siegel-Tukey.

4. When sampling from two populations with different means, but identical symmetric shapes the mean aligned Klotz and Siegel-Tukey tests are reasonably robust with both equal and unequal cell frequencies. The mean aligned Klotz test is more powerful than the corresponding Siegel-Tukey test. In addition it has power equal to or greater than the powers of the Brown-Forsythe and

O'Brien tests. The power of the mean aligned Klotz test does not seem to be affected by the magnitude of the mean difference, and with small effect sizes can have a substantial power advantage relative to the unaligned Klotz test. This suggests that with symmetric distributions, the mean aligned Klotz test can be used regardless of whether there are between group mean differences. Additional research is required to substantiate this conjecture. Fligner (1979) presented a class of distribution-free tests for scale which includes the Siegel-Tukey test. He investigated the effect of small, between group, median differences on the behavior of several tests in the class. The results showed that Type I error rates for various members of the class depended on tailweight of the parent distribution. Consequently, Fligner proposed an adaptive test based on a measure of tailweight. In the adaptive test, the Siegel-Tukey test is used with heavy tailed distributions. Our power results point to the use of the mean aligned Klotz test rather than the Siegel-Tukey test. Moreover, because the mean aligned Klotz test is effective with normal and light tailed distributions, it may be worthwhile to compare it to the tests favored by Fligner for medium and light tailed distributions.

5. When the parent distributions are slightly skewed (skewness = .50) the median aligned Klotz test is reasonably robust for equal and unequal cell frequencies. For either equal, unequal cell frequencies, or both, the other rank tests of scale are not robust. In addition the median aligned Klotz test is as powerful or more powerful than the Brown-Forsythe and O'Brien tests. When the distributions are moderately skewed (skewness = .75) or skewed and leptokurtic (skewness = 1.75, kurtosis = 3.75) none of the rank tests of scale are robust. The Brown-Forsythe test has power equal to or greater than the O'Brien test. These results suggest the need for

research on the efficacy of choosing a procedure based on a measure of skewness.

6. Results for a frequency configuration of 23/17 (inverse relationship between cell frequencies and variances) indicate that the t-test can be liberal with a variance ratio as small as 1:2.5. The power of the scale tests to detect this ratio is quite limited for many of the conditions investigated. For a cell frequency configuration of 15/10 the test is more liberal and the scale tests are less powerful. This suggests that the scale tests are not particularly useful as tests for violations of homoscedasticity. Moreover the Welch-James (1951) procedure does not assume variance equality and has power equivalent to the F-test when the homogeneity assumption is met. The Type I error rate of the Welch-James test is unaffected by variance heterogeneity. When there is an inverse relationship between cell variances and cell frequencies, ANOVA tends to be liberal. Therefore the Welch-James procedure is more appropriate. When there is a direct relationship, ANOVA tends to be conservative and therefore should not be rejected automatically. However the Welch-James test is more powerful and therefore is the procedure of choice. This suggests that the Welch-James test for mean differences should be uniformly adopted when cell frequencies are unequal.

7. The relatively limited power observed with many of the variance ratios suggests the need for total samples larger than 40 when the purpose of the experiment is to test for inequality of variances.

The generality of the conclusions 1 to 6 is limited by our choice of number of treatment groups, total sample sizes, differences in cell frequencies, and identical shapes across treatment groups. The effect of variation in these factors should be investigated.

References

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society*, *A901*, 160, 268-282.

Bartlett, M. S., & Kendall, D.G. (1946). The statistical analysis of variance heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society*, *8*, 128-138.

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

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

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

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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

Fligner, M. AM. (1979). A class of two-sample distribution-free tests for scale. *Journal of the American Statistical Association*, *74*, 889-993.

Games, P. A. (1978). A three-factor model encompassing many possible statistical tests on independent groups. *Psychological Bulletin*, *85*, 168-182.

Games, P. A., Keselman, H. J., & Clinch, J. J. (1979). Tests for homogeneity of variance in factorial designs. *Psychological Bulletin*, *86*, 978-984.

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

Glass, G. V., Peckham, P. D., & Sanders, S. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

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

Klotz, J. (1962). Nonparametric tests for scale. Annals of Mathematical Statistics, 33, 495-512.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, 583-621.

Layard, M. W. J. (1973). Robust large-sample tests of homogeneity of variances. Journal of the American Statistical Association, 68, 195-198.

Lehmann, E. L. (1975). Nonparametrics: Statistical methods based on ranks. San Francisco: Holden-Day.

Levene, H. (1960). Robust test for equality of variances. In I. Olkin (Ed.), Contributions to probability and statistics (pp. 278-292). Stanford , CA: Stanford University Press.

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

Marascuilo, L. A., & McSweeney, M. (1977). Nonparametric and distribution-free methods for the social sciences. Monterey, CA: Brooke/Cole.

Martin, C. G. (1976). Comment on Levy's "An empirical comparison of the Z-variance and Box-Scheffé tests for homogeneity of variance". Psychometrika, 41, 551-556.

Martin, C. G. & Games, P. A. *Selection of subsample sizes for the Bartlett and Kendall test of homogeneity of variance*. Paper presented at the meeting of the American Educational Research Association, Washington, D.C., April 1975 (ERIC document reproduction service MO ED 117 150).

Martin, C. G., & Games, P. A. (1977). ANOVA tests for homogeneity of variance: Non-normality and unequal samples. *Journal of Educational Statistics*, *2*, 187-206.

Miller, R. G. (1968). Jackknifing variances. *Annals of Mathematical Statistics*, *39*, 567-582.

Moses, L. E. (1963). Rank tests of dispersion. *Annals of Mathematical Statistics*, *34*, 973-983.

O'Brien, R. G. (1978). Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika*, *43*, 327-342.

O'Brien, R. B. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, *74*, 877-880.

O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin*, *89*, 570-574.

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

Puri, M. L. (1965). On some tests of homogeneity of variances. *Annals of the Institute of Statistical Mathematics*, *17*, 323-330.

Ramsey, P. H. (1980). Exact Type I error rates for robustness of Student's *t* test with unequal variance. *Journal of Educational Statistics*, *5*, 337-350.

Siegel, S., & Tukey, J. W. (1960). A nonparametric sum of ranks procedure for relative spread in unpaired samples. *American Statistical Association Journal*, *55*, 429-445.

Notes

1. Puri's generalization to k samples does not include the weighting factor $(\frac{N-n_k}{N})$ suggested by Kruskal and Wallis to improve the approximation to the chi-square distribution.

2. The test statistic used here [(N-1) SSB/SST] is a modification of Puri's statistic to include Kruskal and Wallis' adjustment factor.

Table 1

Estimated Type I Error Rates for Independent Samples and Test

		Distributions				
n_1/n_2	$\sigma_1^2:\sigma_2^2$	Normal	Platykurtic	Moderately Skewed	Leptokurtic	Skewed/ Leptokurtic
10/15	1:1.0	.046	.058	.043	.051	.045
	1:1.5	.049	.040	.039	.027	.040
	1:2.0	.041	.028	.045	.042	.048
	1:2.5	.040	.024	.049	.028	.049
	1:3.0	.034	.040	.035	.028	.043
	1:3.5	.028	.027	.035	.025	.039
	1:4.0	.030	.035	.027	.022	.049
15/10	1:1.5	.069	.063	.059	.068	.070
	1:2.0	.069	.061	.078	.053	.062
	1:2.5	.067	.079	.079	.070	.077
	1:3.0	.089	.076	.089	.073	.085
	1:3.5	.085	.077	.075	.088	.096
	1:4.0	.084	.087	.088	.080	.100
23/17	1:1.0	.064	.051	.052	.056	.040
	1:1.5	.051	.056	.057	.057	.055
	1:2.0	.045	.056	.058	.067	.061
	1:2.5	.068	.073	.065	.065	.073
	1:3.0	.060	.066	.065	.072	.079
	1:3.5	.078	.081	.085	.086	.095
	1:4.0	.068	.078	.071	.082	.089
17/23	1:1.5	.038	.041	.045	.047	.047
	1:2.0	.044	.040	.043	.035	.052
	1:2.5	.036	.039	.047	.047	.036
	1:3.0	.037	.038	.036	.040	.035
	1:3.5	.034	.041	.047	.029	.038
	1:4.0	.034	.033	.040	.039	.047
20/20	1:1.0	.044	.056	.047	.043	.048
	1:1.5	.045	.044	.053	.052	.059
	1:2.0	.057	.048	.041	.045	.047
	1:2.5	.049	.055	.047	.053	.069
	1:3.0	.052	.059	.049	.051	.060
	1:3.5	.046	.051	.050	.047	.061
	1:4.0	.071	.053	.056	.039	.062

Note: Nominal alpha level = .05; each figure calculated from 1000 replications.

Table 2

Frequency Distributions and Descriptive Statistics for Six Distributions

	Distributions					
				Slight	Moderate	Skewed/
Interval	Normal	Platykurtic	Leptokurtic	Skew	Skew	Leptokurtic
- ∞ ,-3.0	17		151			
-3.0,-2.5	85		119			
-2.5,-2.0	332		301			
-2.0,-1.5	889	1552	601	882		
-1.5,-1.0	1885	2297	1257	2469	3605	
-1.0,-0.5	2470	2917	2816	3516	3976	8555
-0.5, 0.0	3826	3235	4745	3851	3591	4219
0.0, 0.5	3817	3177	4753	3474	2053	2577
0.5, 1.0	3038	2805	2748	2590	2345	1777
1.0, 1.5	1849	2411	1343	1626	1552	1142
1.5, 2.0	855	1606	586	888	1039	671
2.0, 2.5	332		263	456	520	440
2.5, 3.0	86		178	171	230	268
3.0, ∞	19		139	77	89	351
Mean	-.0015	.0049	.0004	-.0053	.0009	- .0063
Variance	.9836	1.0109	1.0292	.9887	1.0631	.9774
Skewness	.0004	- .0005	- .1297	.5044	.7266	1.6820
Kurtosis	-.0938	-1.0131	3.5547	-.0216	- .0846	3.1517

Note: Results based on 20,000 observations.

Table 3

Estimated Actual Type I Error Rates for Tests on Variance

		Distributions				
n_1/n_2	Test[a]	Normal	Platykurtic	Moderate Skew	Leptokurtic	Skewed Leptokurtic
10/15	OB	.046	.046	.051	.032[t]	.064
	BF	.045	.038	.033[t]	.033[t]	.049
	ST	.066*	.045	.051	.041	.046
	K	.053	.046	.044	.037	.047
0/20	OB	.051	.051	.050	.033[t]	.064
	BF	.045	.031[t]	.046	.041	.058
	ST	.052	.053	.053	.053	.050
	K	.056	.049	.042	.050	.043
17/23	OB	.054	.059	.059	.035[t]	.050
	BF	.047	.043	.040	.040	.049
	ST	.053	.057	.050	.048	.054
	K	.050	.044	.042	.045	.054

t - indicates a conservative test
* - indicates a liberal test
a - OB=O'Brien, BF=Brown-Forsythe, ST=Siegel-Tukey, K=Klotz

Table 4

Estimated Power for the Tests on Variance

		20/20				17/23				23/17			
Distribution	$\sigma_1^2:\sigma_2^2$	OB[a]	BF	ST	K	OB	BF	ST	K	OB	BF	ST	K
Normal	1:1.5	.119	.104	.104	.125	.087	.094	.085	.102	.138	.099	.105	.136
	1:2.0	.258	.244	.201	.250	.200	.213	.200	.220	.262	.199	.198	.262
	1:2.5	.391	.366	.321	.393	.310	.352	.276	.352	.452	.361	.317	.437
	1:3.0	.526	.490	.407	.533	.460	.496	.419	.481	.573	.504	.426	.548
	1:3.5	.611	.610	.501	.625	.567	.588	.494	.596	.703	.627	.536	.677
	1:4.0	.704	.696	.583	.700	.673	.711	.600	.710	.758	.695	.585	.753
Platykurtic	1:1.5	.180	.121	.136	.214	.137	.101	.112	.159	.184	.088	.119	.225
	1:2.0	.409	.289	.266	.442	.344	.254	.236	.358	.488	.288	.298	.494
	1:2.5	.666	.482	.435	.657	.577	.454	.399	.575	.645	.451	.428	.621
	1:3.0	.775	.618	.515	.742	.738	.589	.511	.727	.801	.604	.530	.770
	1:3.5	.861	.726	.601	.828	.846	.761	.649	.815	.872	.724	.612	.838
	1:4.0	.918	.841	.705	.893	.880	.796	.685	.857	.408	.787	.685	.889
Moderate Skew	1:1.5	.140	.113	.152	.221	.113	.101	.136	.204	.127	.080	.148	.230
	1:2.0	.276	.236	.326	.467	.236	.228	.308	.405	.303	.226	.317	.472
	1:2.5	.399	.369	.441	.639	.332	.335	.419	.549	.453	.344	.452	.646
	1:3.0	.553	.517	.571	.748	.474	.474	.555	.702	.593	.497	.580	.761
	1:3.5	.617	.582	.638	.806	.561	.574	.614	.746	.695	.616	.667	.846
	1:4.0	.738	.705	.726	.861	.629	.669	.716	.823	.760	.679	.718	.867
Leptokurtic	1:1.5	.058	.090	.088	.087	.046	.067	.073	.075	.104	.102	.105	.102
	1:2.0	.134	.167	.156	.169	.087	.146	.146	.146	.141	.157	.160	.180
	1:2.5	.197	.261	.231	.262	.131	.240	.204	.208	.281	.305	.220	.311
	1:3.0	.266	.367	.327	.370	.192	.329	.331	.335	.328	.390	.339	.400
	1:3.5	.346	.452	.396	.459	.251	.423	.374	.427	.432	.494	.460	.510
	1:4.0	.423	.525	.447	.525	.502	.514	.478	.499	.492	.564	.495	.560
Skewed-Leptokurtic	1:1.5	.067	.077	.256	.367	.065	.079	.234	.326	.099	.083	.272	.403
	1:2.0	.139	.147	.500	.598	.102	.134	.460	.498	.169	.147	.520	.630
	1:2.5	.180	.203	.621	.668	.152	.194	.645	.674	.243	.222	.642	.718
	1:3.0	.239	.282	.739	.780	.194	.254	.746	.756	.295	.308	.740	.816
	1:3.5	.301	.360	.817	.846	.223	.307	.781	.770	.350	.363	.803	.867
	1:4.0	.359	.416	.844	.869	.249	.345	.833	.800	.430	.456	.853	.892

a - OB=O'Brien, BF=Brown-Forsythe, K-Klotz, ST-Siegel-Tukey

Table 5

Power Partial Orders When There Are No Between Group Mean Differences

<table>
<tr><th>Distribution</th><th>Partial Order [a]</th></tr>
<tr><td>Normal</td><td>OB-BF-K
|
ST</td></tr>
<tr><td>Platykurtic</td><td>K-OB
|
BF
|
ST</td></tr>
<tr><td>Skewed</td><td>K
|
ST
|
OB
|
BF</td></tr>
<tr><td>Leptokurtic</td><td>K-BF
|
ST
|
OB</td></tr>
<tr><td>Skewed/Leptokurtic</td><td>K-ST
|
BF
|
OB</td></tr>
</table>

[a] - OB=O'Brien, BF=Brown-Forsythe, K=Klotz, ST=Siegel-Tukey.

Table 6

Estimated Actual Type I Error Rates for Tests on Variance When There Are Between Group Mean Differences

				20/20						23/17			
Distribution	Effect Size	ST[a]	K	STM	KM	STMD	KMD	ST	K	STM	KM	STMD	KMD
Normal	S	.042	.043	.050	.052	.038	.038	.054	.051	.064	.063	.087*	.083
	M	.052	.036	.079	.064	.057	.041	.041	.031	.054	.061	.081*	.047
	L	.031^t	.018^t	.047	.060	.036	.044	.034^t	.018^t	.054	.056	.078*	.042
Platykurtic	S	.050	.049	.051	.052	.039	.029^t	.052	.042	.064	.051	.092*	.031^t
	M	.041	.020^t	.052	.054	.034	.032^t	.053	.019^t	.071*	.055	.092*	.036
	L	.028^t	.007^t	.058	.058	.033^t	.040	.026^t	.005^t	.054	.052	.086*	.039
Leptokurtic	S	.052	.056	.064	.061	.044	.051	.060	.053	.066*	.060	.080*	.046
	M	.042	.034^t	.054	.051	.050	.048	.037	.039	.060	.055	.077*	.039
	L	.038	.019^t	.070*	.070*	.061	.063	.040	.033^t	.051	.037	.081*	.035^t
Slight Skew	S	.052	.046	.064*	.079*	.054	.054	.047	.044	.057	.069*	.078*	.040
	M	.055	.054	.071*	.071*	.048	.067*	.047	.030^t	.074*	.077*	.098*	.048
	L	.047	.031^t	.079*	.081*	.052	.049	.033^t	.029^t	.063	.082*	.084*	.065*
Moderate Skew	S	.068*	.073*	.118*	.140*	.077*	.095*	.061	.060	.096*	.132*	.115*	.084*
	M	.076*	.067*	.108*	.133*	.072*	.088*	.086*	.069*	.105*	.140*	.136*	.091*
	L	.103*	.063	.086*	.119*	.058	.075*	.090*	.047	.114*	.134*	.139*	.084*
Skewed-Leptokurtic	S	.210*	.216*	.350*	.419*	.233*	.274*	.198*	.201*	.374*	.413*	.343*	.255*
	M	.422*	.270*	.329*	.410*	.219*	.246*	.379*	.206*	.346*	.411*	.328*	.248*
	L	.496*	.205*	.347*	.396*	.243*	.272*	.369*	.134*	.366*	.439*	.321*	.264*

[a] ST=Siegel-Tukey, K=Klotz, STM=Siegel-Tukey with adjustment for sample mean, KM=Klotz with adjustment for sample mean, KMD=Klotz with adjustment for sample median, STMD=Siegel-Tukey with adjustment for sample median.

t - indicates a conservative estimated Type I error rate.
* - indicates a liberal estimated Type I error rate.

Table 7

Estimated Powers for Tests on Variance[a,b]

Distribution	$\sigma_1^2:\sigma_2^2$	Test Statistic[d]							
		OB	BF	ST[c]	K[c]	STM	KM	STMD	KMD
Normal	1:1.5	.106	.099	.089	.095	.107	.137	.086	.106
	1:2.0	.249	.231	.172	.236	.211	.295	.176	.241
	1:2.5	.404	.374	.294	.386	.326	.449	.274	.382
	1:3.0	.529	.510	.399	.509	.438	.606	.376	.539
	1:3.5	.653	.635	.506	.649	.556	.723	.483	.649
	1:4.0	.739	.728	.587	.699	.639	.805	.565	.745
Platykurtic	1:1.5	.170	.106	.118	.174	.126	.207	.095	.141
	1:2.0	.423	.278	.251	.400	.272	.468	.199	.356
	1:2.5	.650	.489	.369	.584	.433	.667	.327	.554
	1:3.0	.775	.617	.552	.732	.543	.783	.462	.690
	1:3.5	.864	.747	.619	.831	.633	.866	.535	.780
	1:4.0	.921	.828	.672	.875	.716	.923	.625	.851
Leptokurtic	1:1.5	.062	.073	.071	.072	.089	.105	.068	.083
	1:2.0	.128	.169	.161	.175	.238	.209	.142	.170
	1:2.5	.208	.266	.250	.265	.264	.308	.220	.275
	1:3.0	.271	.376	.304	.341	.359	.424	.308	.376
	1:3.5	.340	.469	.385	.435	.444	.509	.384	.458
	1:4.0	.413	.540	.486	.535	.519	.594	.453	.538
Slight Skew	1:1.5	.121	.105	.095	.107			.098	.129
	1:2.0	.256	.236	.192	.265			.195	.280
	1:2.5	.414	.375	.306	.428			.288	.426
	1:3.0	.554	.521	.413	.571			.408	.564
	1:3.5	.647	.623	.520	.681			.488	.659
	1:4.0	.739	.729	.601	.751			.579	.758
Moderate Skew	1:1.5	.117	.090						
	1:2.0	.254	.225						
	1:2.5	.423	.386						
	1:3.0	.542	.488						
	1:3.5	.642	.615						
	1:4.0	.703	.683						
Skewed-Leptokurtic	1:1.5	.080	.083						
	1:2.0	.143	.138						
	1:2.5	.189	.209						
	1:3.0	.243	.278						
	1:3.5	.291	.340						
	1:4.0	.343	.396						

a - Results refer to the 20/20 cell frequencies.

b - Power figures are not reported for tests that were liberal.

c - Results refer to conditions with small mean effect sizes.

d - OB=O'Brien, BF=Brown-Forsythe, ST=Siegel-Tukey,K= Klotz, STM=Siegel-Tukey with adjustment for sample mean, KM=Klotz with adjustment for sample mean, STMD =Siegel-Tukey with adjustment for sample median, KMD=Klotz with adjustment for sample median.

Table 8

Estimated Powers For Tests on Variance[a,b]

		Test Statistic [d]							
Distribution	$\sigma_1^2:\sigma_2^2$	OB	BF	ST [c]	K [c]	STM	KM	STMD	KMD
Normal	1:1.5	.128	.095	.097	.122	.107	.142		.106
	1:2.0	.288	.226	.222	.281	.217	.307		.245
	1:2.5	.450	.361	.313	.409	.334	.461		.385
	1:3.0	.581	.502	.423	.570	.427	.591		.516
	1:3.5	.695	.610	.516	.657	.422	.700		.630
	1:4.0	.760	.684	.580	.715	.588	.769		.702
Platykurtic	1:1.5	.196	.101	.124	.203	.130	.212		.145
	1:2.0	.445	.269	.280	.448	.266	.459		.342
	1:2.5	.666	.448	.397	.642	.410	.657		.539
	1:3.0	.802	.611	.538	.761	.538	.790		.679
	1:3.5	.883	.726	.603	.819	.639	.872		.770
	1:4.0	.933	.817	.704	.901	.704	.919		.850
Leptokurtic	1:1.5	.081	.083	.091	.098	.105	.119		.094
	1:2.0	.161	.166	.167	.180	.185	.223		.183
	1:2.5	.264	.291	.256	.302	.286	.351		.291
	1:3.0	.345	.361	.323	.358	.371	.423		.365
	1:3.5	.429	.479	.417	.494	.435	.524		.458
	1:4.0	.476	.542	.470	.541	.498	.591		.527
Slight Skew	1:1.5	.132	.100	.084	.017				.127
	1:2.0	.280	.225	.193	.226				.272
	1:2.5	.411	.353	.263	.352				.407
	1:3.0	.556	.505	.422	.523				.545
	1:3.5	.656	.615	.490	.617				.658
	1:4.0	.728	.696	.568	.713				.727
Moderate Skew	1:1.5	.155	.101	.095	.148				
	1:2.0	.292	.208	.213	.368				
	1:2.5	.447	.362	.391	.608				
	1:3.0	.572	.474	.473	.701				
	1:3.5	.684	.602	.595	.808				
	1:4.0	.743	.685	.650	.848				
Skewed-Leptokurtic	1:1.5	.099	.081						
	1:2.0	.182	.154						
	1:2.5	.227	.214						
	1:3.0	.311	.300						
	1:3.5	.354	.350						
	1:4.0	.408	.427						

a - Results refer to the 23/17 cell frequencies.

b - Power figures are not reported for tests that were liberal.

c - Results refer to conditions with small mean effect sizes.

d - OB = O'Brien, BF=Brown-Forsythe, ST=Siegel-Tukey, K=Klotz, STM = Siegel-Tukey with adjustment for sample mean, KM=Klotz with adjustment for sample mean, STMD-Siegel-Tukey with adjustment for sample median, KMD=Klotz with adjustment for sample median.

Table 9

Estimated Power for Tests on Variance[a,b]

Distribution	$\sigma_1^2:\sigma_2^2$	Test Statistic[d] OB	BF	ST[c]	K[c]	STM	KM	STMD	KMD
Normal	1:1.5	.098	.095	.091	.093	.108	.125		.102
	1:2.0	.216	.217	.189	.213	.210	.269		.219
	1:2.5	.349	.361	.285	.365	.335	.436		.368
	1:3.0	.455	.410	.374	.440	.454	.554		.469
	1:3.5	.559	.590	.487	.563	.536	.657		.590
	1:4.0	.640	.690	.545	.640	.617	.749		.675
Platykurtic	1:1.5	.159	.112	.122	.170	.131	.185		.128
	1:2.0	.384	.288	.246	.353	.282	.428		.326
	1:2.5	.586	.443	.377	.550	.423	.610		.501
	1:3.0	.734	.602	.485	.678	.540	.764		.660
	1:3.5	.835	.723	.600	.781	.646	.848		.754
	1:4.0	.894	.806	.676	.853	.728	.902		.829
Leptokurtic	1:1.5	.055	.077	.078	.069	.092	.106		.087
	1:2.0	.091	.160	.165	.165	.182	.206		.174
	1:2.5	.135	.237	.221	.238	.264	.287		.252
	1:3.0	.207	.343	.206	.318	.344	.403		.341
	1:3.5	.256	.433	.377	.407	.441	.470		.424
	1:4.0	.296	.488	.421	.451	..497	.566		.506
Slight Skew	1:1.5	.096	.091	.085	.088				.105
	1:2.0	.221	.218	.185	.238				.241
	1:2.5	.361	.352	.313	.398				.396
	1:3.0	.481	.496	.425	.540				.532
	1:3.5	.580	.600	.511	.648				.632
	1:4.0	.660	.692	.591	.729				.728
Moderate Skew	1:1.5		.098	.085					
	1:2.0		.216	.215					
	1:2.5		.350	.336					
	1:3.0		.473	.452					
	1:3.5		.574	.536					
	1:4.0		.674	.654					
Skewed-Leptokurtic	1:1.5	.073	.077						
	1:2.0	.119	.133						
	1:2.5	.156	.169						
	1:3.0	.190	.237						
	1:3.5	.234	.307						
	1:4.0	.264	.380						

a - Results refer to the 17/23 cell frequencies.

b- Power figures not reported for tests that were liberal.

c- Results refer to the conditions with the small effect sizes.

d- OB =O'Brien, BF=Brown-Forsythe, ST=Siegel-Tukey, K=Klotz, STM= Siegel-Tukey with adjustment for sample mean, KM=Klotz with adjustment for sample mean, KMD=Klotz with adjustment for sample median, STMD=Siegel-Tukey with adjustment for sample median.

Table 10

Power Partial Orders When There Are Between Group Mean Differences

	Cell Frequencies		
Distribution	20/20	23/17	17/23
Normal	KM[a] | OB——BF	KM——OB | BF	KM | BF | OB
Platykurtic	KM——OB | BF	KM——OB | BF	KM | OB | BF
Leptokurtic	KM | BF | OB	KM | BF | OB	KM | BF | OB

a - OB=O'Brien, BF=Brown-Forsythe, KM=Klotz with mean aligned data.