Florida Journal of Educational Research
Fall 1985, Vol. 27, No. 1, Pp. 61-82

Type I Error Rate and Power of Rank Transform ANOVA When Populations are Non-Normal and Have Equal Variance

Stephen F. Olejnik
University of Georgia

and

James Algina
University of Florida

ABSTRACT. The rank transformation approach to analysis of variance as a solution to the Behrens-Fisher problem is examined. Using simulation methodology four parameters were manipulated for the two group design: (1) ratio of population variances; (2) distribution form; (3) sample size and (4) population mean difference. The results indicated that while the rank transform approach was less sensitive to variance inequality than the parametric ANOVA F-ratio, unacceptably high Type I error rates were obtained when cell frequencies and group variances were inversely related. With equal cell frequencies and/or when cell frequencies were directly related to group variances, appropriate Type I error rates were obtained. Under these conditions however, the Brown-Forsythe procedure for comparing, group means provided greater power except when the sampled distribution was leptokurtic.

Both empirical and analytic studies have repeatedly shown that parametric analysis procedures for comparing group means are extremely sensitive to population variance inequality when sample sizes are markedly unequal. When sample size and group variance

are positively correlated, the nominal significance level is underestimated while with a negative relationship between sample size and group variance, the nominal significance level is overestimated (Glass, Peckham and Sanders, 1972). Even with equal sample sizes, Ramsey (1980) has shown that the actual probability of a Type I error for the t-test may either over or underestimate the nominal significance level. Developing alternative data analysis strategies when population variances differ, also known as the Behrens-Fisher problem, has therefore been an area of considerable interest, and several solutions to the problem have been suggested. The procedure offered by Welch (1947, 1951), in particular, has gained considerable attention. Welch's solution modifies the ANOVA F-ratio by weighting the sample means by the ratio of the group frequency to the group variance. In addition the degrees of freedom error are adjusted so that the computed statistic approximates the F distribution. Wang (1971) has shown that this approximation is satisfactory for most situations. James (1951) suggested a similar weighting procedure but used the chi-square distribution as the reference distribution.

Recently Brown and Forsythe (1974a) have suggested a slightly different approach to the Behrens-Fisher problem. Their statistic takes the ratio of the sums of squares between groups to a weighted sum of group variances. The test statistic has an approximated F distribution. For the two group case, the Welch and the Brown-Forsythe procedures are identical (Brown-Forsythe, 1974a), but differ when multiple groups are compared. Both procedures have been generalized for factorial designs (Brown and Forsythe, 1974b; Johansen, 1980; Algina and Olejnik, 1984).

A number of investigations have studied both of these strategies and compared them with respect to their Type I error rates and statistical power. The results of these studies have shown that both approaches are insensitive to variance inequality when the sampled distributions were normal (Kohr and Games, 1974; Brown and Forsythe, 1974a; Levy, 1978; Dijkstra and Werter, 1981; Lee and Fung, 1983). Under non-normal parent distributions the Brown-Forsythe

statistic was shown to provide appropriate Type I error rates (Clinch and Keselman, 1982; Lee and Fung, 1983). The results with Welch's approach however have been mixed and inconsistent. There is some evidence to indicate that the approach is liberal for skewed distributions when the number of levels of the grouping factor is four or more (Clinch and Keselman, 1982; Levy, 1978). Levy on the other hand found appropriate Type I error rates when there were three levels of the independent variable and the sampled population had a chi-square distribution. When data were sampled from heavy-tailed distributions, some results have indicated that Welch's procedure provides a conservative test of group means (Yuen, 1974; Lee and Fung, 1983). Other evidence however indicates appropriate Type I error rates (Clinch and Keselman, 1982). Differences in these conclusions may be a function of the degree to which the sampled populations departed from normality. Finally for light-tailed distributions, the Welch procedure has been shown to provide a liberal test of means (Levy, 1978), but other results indicate that appropriate Type I error rates are possible (Yuen, 1974).

When the procedures were compared in terms of their statistical power, the results have been mixed but generally consistent. Both procedures provide comparable power when the population distributions are normal and the variances are equal. Under this condition both procedures are only slightly less powerful than the ANOVA F-ratio (Brown and Forsythe, 1974a; Dijkstra and Werter, 1981; Clinch and Keselman, 1982; Lee and Fung, 1983). With unequal variances and a normal parent distribution, the Brown-Forsythe approach provided greater power when the extreme mean had lower variance, while the Welch procedure was more sensitive if the extreme mean had the large variance (Brown and Forsythe, 1974a; Dijkstra and Werter, 1981; Lee and Fung, 1983). Clinch and Keselman however found very little difference in statistical power between the procedures for this condition. The statistical power for all of the procedures studied however was relatively low, and that may explain the inconsistency in the results reported by Clinch and Keselman. Finally for heavy-tailed distributions the

Welch procedure provided a slight power advantage (Clinch and Keselman, 1982; Lee and Fung, 1983).

Recently, Dauphin (1983) considered a different approach to the Behrens-Fisher problem. She suggested transforming the original data by using ranks before group means are compared. After ranking the data from highest to lowest across all comparison groups, the parametric analysis of variance F-ratio is computed. This strategy of transforming data using ranks before computing parametric analyses has been suggested by Conover and Iman (1981) as a linking procedure between parametric and nonparametric analysis strategies. They have suggested that the ranking approach can be used in a variety of research contexts and considerable research has been conducted using this approach generally with positive results. Nath and Duran (1981a, 1981b) studied the procedure when two group means are to be compared, Conover and Iman (1982) applied the approach to analysis of covariance, Iman and Conover (1979) used the rank transformation in a regression problem, and Iman (1974) studied the approach for factorial designs when an interaction was present. Although the theoretical rationale for the procedure is not fully developed, progress in that direction has been reported by Iman, Hora and Conover (1984).

The use of the rank transformation has been motivated primarily as an alternative analysis strategy to parametric statistics when sampled distributions were non-normal. In this context the rank transformation has often provided a more sensitive test of the location parameter than the parametric alternative. The rationale of applying the rank transformation as a solution to the Behrens-Fisher problem was based on previous findings that nonparametric strategies, while affected by variance inequality, are less sensitive than the parametric alternatives (Wetherill, 1960). Glazer (1963), for example, empirically demonstrated Wetherill's asymptotic results showing that the Wilcoxon-Mann-Whitney probability of a Type I error was less affected than Student's t-test for independent sample means when the population variances differed. Since the rank transform is monotonically

related to the Wilcoxon test, Dauphin expected similar conclusions. Her results confirmed her expectation showing that the actual Type I error rate for the rank transform did not deviate greatly from the nominal significance level when the sampled population was normal.

Since the rank transformation has gained considerable interest and Dauphin's results indicate that the approach may have some merit in some situations, it was decided to examine this proposed solution to the Behrens-Fisher problem a little closer. Specifically the purpose of the study was to analyze the empirical Type I error rates of the rank transform ANOVA with parametric analysis of variance and Brown-Forsythe's procedure when population variances differed, and the distributions were normal or non-normal. In addition, for those situations where appropriate Type I error rates were observed, the statistical power estimates for small, medium and large differences in group means were compared.

Computer Simulation

In order to calculate empirical Type I error rates and statistical power estimates for each of the competing analysis strategies under a variety of conditions, four factors were manipulated: 1) sample size; 2) distribution form; 3) population mean difference and 4) population variance inequality. Although all three of the procedures can be used for comparing group means in multiple group designs, including factorial designs, the present investigation was limited to comparisons between two groups.

Sample Size. Samples of (10,15), (15,10), (20,20), (17,23), and (23,17) were included in the investigation. The sample sizes considered here were thought to be moderate and representative of those often found in research studies in the social sciences. Small departures from equal n's were chosen to represent common attrition rates in social research.

Distribution Form. A normal and four non-normal parent distributions were considered. The non-normal distributions included a light-tailed, platykurtic

distribution, a symmetric, leptokurtic (heavy-tailed) distribution, 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.

Population Mean Difference. To study the Type I error rates of the three procedures, data were generated from populations which had a common mean. Power estimates were obtained by comparing the proportion of hypotheses rejected when data were sampled from populations which differed by .2, .5, or .8 pooled standard deviation units. These differences in the location parameters have been suggested by Cohen (1977) as representing small, medium, and large effects respectively.

Population Variances. The present study considered populations which had common variances as well as six levels of variance inequality. Specifically data were generated from populations with the following variance pairs: (1,1), (1,1.5), (1,2.0), (1,2.5), (1,3.0), (1,3.5), and (1,4.0). The choice of these variance differences was based on two considerations. First, it was believed that the conditions considered reflected common situations encountered by applied researchers. Second, it was believed that with the unequal sample size combinations studied, the variance differences would affect the Type I error rate of the parametric ANOVA F-ratio.

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 varied from 0, .2, .5, or .8 pooled standard deviation units to study the effect population mean difference. In all cases the shift parameter was added to the second group so that $\mu_1 < \mu_2$. The random error component ε_{ij} was generated using the SAS NORMAL function to simulate scores, X_{ij}, from a standard normal distribution. For a normally distributed error component, ε_{ij} was set equal to X_{ij}. For a non-normally distributed component, X_{ij} was transformed using a power function

developed by Fleishman (1978): $\varepsilon_{ij} = [(dX_{ij}+c)X_{ij}+b]X_{ij}+a$. The constants a, b, c and d were chosen to transform the standard normal variable to a variable with known skewness and kurtosis and null mean and unit variance. Four non-normal distributions were considered in the study. Descriptive statistics and frequency distributions at half standard deviation intervals are included in Table 1. Values reported in the table are based on 20,000 random variables generated for each distribution. The variance of the observations in group one was kept constant at 1 for all conditions studied while the variance of the second group was increased from 1 to 4 in increments of .5 units by multiplying the random error component by the desired standard deviation.

Computed Test Statistics. In each sample generated, the group means were compared using the parametric analysis of variance F-ratio, the Brown-Forsythe (1974a) test statistic, and the rank transform analysis of variance F-ratio.

The parametric analysis of variance F-ratio is computed as the ratio of the mean square between group means to the pooled within group variance:

$$F = \frac{\Sigma_j n_j(\bar{Y}._j-\bar{Y}..)^2/(J-1)}{\Sigma_j (n_j-1)S_j^2/(N-J)}$$

where n_j is the number of observations in the j^{th} group; N is the total number of observations in the sample; J is the number of groups in the study; $\bar{Y}._j$ is the sample mean for the j^{th} group; $\bar{Y}..$ is the grand mean; S_j^2 is the variance of the j^{th} group. The critical test statistic is obtained from the F distribution with J-1 and N-J degrees of freedom.

The rank transform ANOVA F-ratio is computed using the same formula as parametric ANOVA with the dependent variable obtained by replacing the original observations with the rank of the observation. The observations are ranked by assigning a 1 to the lowest observation and N to the highest observation in the

Table 1

Frequency Distributions and Descriptive Statistics

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

total sample across all groups. If ties are present the average rank is assigned to all tied observations. The F-ratio is computed as:

$$F_R = \frac{\Sigma_j n_j(\bar{R}_{.j}-\bar{R}_{..})^2/J-1}{\Sigma_j (n_j-1)S^2_{R_j}/N-J}$$

where $\bar{R}_{.j}$ is the mean rank of group j; $\bar{R}_{..}$ is the grand mean rank $N+1/2$; S^2_{Rj} is the variance on the ranks for the j^{th} group. The critical test statistic for the rank transform F-ratio is the same as that for the parametric ANOVA.

The Brown-Forsythe statistic is obtained as the ratio of the sum of squares between groups and a weighted sum of within group variance:

$$F_{BF} = \frac{\Sigma_j n_j(\bar{Y}_{.j}-\bar{Y}_{..})^2}{\Sigma_j (1-\frac{n_j}{N})S^2_j}$$

where the terms are defined as stated above. The Brown-Forsythe F statistic has an F distribution with degrees of freedom J-1 and f where 1/f is equal to

$$[c_j^2/(n_j-1)] \text{ and } c_j=(1-\frac{n_j}{N})S^2_j/\ [\Sigma_j^J (1-\frac{n_j}{N})S^2_j\ ]$$

For each condition, 1000 replications of the three statistics were computed, and the frequency at which each procedure rejected the null hypothesis of equal population means at the .05 level was recorded. 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, observed Type I error rates outside the interval (.036, .064) were considered nonrobust.

Results

The results of the study are reported in two sections. In the first section the empirical Type I error rates for the Brown-Forsythe, the parametric ANOVA, and the rank transform ANOVA are presented for increasing variance inequality with equal and unequal sample size combinations. The second section presents the proportions of hypotheses rejected when population differed by .2, .5, or .8 pooled standard deviation units representing small, medium, and large effect sizes. The power results are reported only for those conditions where appropriate Type I error rates were obtained.

Type I Error Rates

As a preliminary test of the computer program and the data generation procedure, data for three sample size combinations were generated from populations identical in their form, scale, and location. The sample means for these samples were compared using the three analysis procedures under consideration, and the proportion of hypotheses rejected at a nominal significance level of .05 were recorded. Table 2 reports the results of these analyses for the five distribution forms studied. None of the observed proportions exceeded two standard errors above or below the expected five percent level. These results therefore support the adequacy of the data generation and analysis procedures used in the study.

The observed Type I error rates, as the difference in group variances increased, are reported in Table 3 for the five distribution forms with equal and unequal sample size combinations. For the normal distribution the results reported here are consistent with those presented by Dauphin (1983). The Type I error rate for the rank transform ANOVA was affected to a lesser degree than the parametric ANOVA F-ratio. However, for situations where the smaller samples had greater variance, the proportion of Type I errors were more than two standard errors above the nominal significance level and therefore judged as being unacceptably high. When large samples had greater

Table 2

Type I Error Rates for the Brown-Forsythe (BF), Parametric ANOVA (F), and the Rank Transform (RF) ANOVA

	Sample Size								
	15/10			20/20			23/17		
Distribution	BF	F	RF	BF	F	RF	BF	F	RF
Normal	.044	.046	.050	.044	.044	.043	.060	.064	.059
Platykurtic	.057	.058	.058	.056	.056	.052	.054	.051	.048
Skewed	.050	.043	.050	.052	.052	.055	.055	.057	.054
Leptokurtic	.048	.051	.056	.041	.043	.046	.058	.056	.061
Skewed and Leptokurtic	.045	.045	.053	.046	.048	.052	.044	.040	.045

Note: Nominal $\alpha = .05$.

Table 3

Type 1 Error Rates for the Brown-Forsythe (BF), Parametric ANOVA (F), and the Rank Transform ANOVA (RF)

Distribution	Variance Ratio $\sigma_1^2:\sigma_2^2$	Sample Size (n_1/n_2) 10/15			15/10			20/20			17/23			23/17		
		BF	F	RF	BF	F	RF	BF	F	RF	BF	F	RF	BF	F	RF
Normal	1:1.5	.059	.049	.052	.050	.069	.066	.044	.045	.042	.047	.038	.050	.046	.051	.052
	1:2.0	.055	.041	.055	.055	.069	.066	.054	.057	.066	.057	.044	.042	.037	.045	.046
	1:2.5	.056	.040	.043	.048	.067	.068	.049	.049	.056	.046	.036	.047	.055	.068	.069
	1:3.0	.055	.034	.045	.053	.089	.079	.050	.052	.061	.044	.037	.045	.042	.060	.056
	1:3.5	.049	.028	.037	.055	.085	.075	.046	.046	.049	.051	.034	.040	.063	.078	.074
	1:4.0	.044	.030	.041	.043	.084	.081	.067	.071	.080	.047	.034	.047	.043	.068	.067
Platykurtic	1:1.5	.043	.040	.047	.060	.063	.068	.043	.043	.043	.044	.041	.048	.048	.056	.055
	1:2.0	.040	.028	.034	.050	.061	.066	.048	.048	.054	.045	.040	.044	.048	.056	.060
	1:2.5	.039	.024	.034	.060	.079	.078	.053	.055	.058	.053	.039	.047	.058	.073	.067
	1:3.0	.058	.040	.059	.052	.076	.075	.059	.059	.062	.055	.038	.046	.045	.066	.072
	1:3.5	.045	.027	.040	.052	.077	.078	.049	.051	.061	.056	.041	.058	.049	.081	.074
	1:4.0	.055	.035	.045	.059	.087	.094	.052	.053	.063	.050	.033	.051	.053	.078	.077
Skewed	1:1.5	.038	.039	.052	.060	.059	.069	.057	.057	.057	.048	.045	.055	.053	.060	.071
	1:2.0	.057	.045	.054	.054	.078	.090	.041	.041	.056	.055	.043	.060	.051	.057	.075
	1:2.5	.054	.049	.071	.065	.079	.089	.046	.047	.073	.057	.047	.072	.053	.072	.093
	1:3.0	.051	.035	.052	.057	.089	.088	.048	.049	.080	.047	.036	.065	.053	.070	.081
	1:3.5	.049	.035	.064	.048	.075	.098	.049	.050	.079	.063	.047	.086	.041	.068	.097
	1:4.0	.045	.027	.053	.054	.088	.109	.053	.056	.079	.057	.040	.081	.059	.077	.110

Leptokurtic	1:1.5	.034	.027	.038	.049	.068	.061	.052	.052	.049	.053	.047	.054	.045	.057	.049
	1:2.0	.053	.042	.050	.036	.053	.045	.043	.045	.053	.047	.035	.050	.053	.067	.062
	1:2.5	.046	.028	.043	.039	.070	.063	.051	.053	.048	.064	.047	.053	.049	.065	.053
	1:3.0	.046	.028	.038	.041	.073	.052	.050	.051	.057	.049	.040	.054	.050	.072	.069
	1:3.5	.036	.025	.037	.051	.088	.077	.045	.047	.048	.044	.029	.051	.061	.086	.071
	1:4.0	.040	.022	.048	.041	.080	.073	.037	.039	.042	.064	.039	.055	.051	.082	.079
Skewed and Leptokurtic	1:1.5	.039	.040	.067	.068	.070	.103	.058	.059	.108	.053	.047	.089	.046	.055	.097
	1:2.0	.046	.048	.095	.056	.062	.120	.043	.047	.119	.056	.052	.145	.055	.061	.149
	1:2.5	.050	.049	.116	.067	.077	.147	.068	.069	.160	.039	.036	.137	.068	.073	.182
	1:3.0	.045	.043	.118	.071	.085	.150	.060	.060	.190	.042	.035	.155	.071	.079	.188
	1:3.5	.045	.039	.110	.077	.096	.170	.058	.061	.192	.045	.038	.203	.077	.095	.209
	1:4.0	.056	.049	.149	.076	.100	.175	.059	.062	.229	.066	.047	.192	.068	.089	.221

Note: Nominal $\alpha = .05$

variance the rank transform had acceptable Type I error rates while the ANOVA F-ratio underestimated the nominal significance level. With equal sample sizes, both the ANOVA F and the rank transform were not seriously affected by variance inequality. The Brown-Forsythe procedure provided appropriate Type I error rates for all degrees of variance inequality and sample size combinations.

With symmetric, non-normal distributions the observed Type I error rates were similar to those obtained under the normal populations. The rank transform ANOVA had Type I error rates which were affected to a lesser degree than the parametric ANOVA F-ratio. Error rates within the acceptable range were obtained for the rank transform approach when sample sizes were equal and when the larger sample size had greater variance. When the sample with fewer observations had greater variance, the observed Type I error rate exceeded the nominal significance level by more than two standard errors. When samples were selected from skewed populations, the rank transform approach had observed Type I error rates overestimating the nominal significance level for all sample size combinations except when samples of (10,15) were selected. Under the latter condition, appropriate Type I error rates were obtained. The Type I error rates for the ANOVA F-ratio were as expected and were similar to those obtained under normal distributions. With the skewed and leptokurtic distribution, the rank transform became quite liberal, even for the condition with sample sizes of (10,15). Again Type I error rates for the parametric F were similar to those obtained with the normal distribution. The Brown-Forsythe procedure overestimated the nominal significance level when the samples distribution was both skewed and leptokurtic and the larger variance was matched with the samples having fewer observations. When larger samples were matched with lower variance, appropriate Type I error rates were obtained. These results were consistent with those reported by Clinch and Keselman (1982) in their analysis of Welch's procedure.

In summarizing these results, the observed Type I error rates for the rank transform was affected to a

lesser degree than the parametric ANOVA F-ratio when the samples distributions were symmetric. This conclusion is consistent with that predicted by Wetherill (1960) and previously demonstrated for the normal distribution by Dauphin (1983). With skewed distributions however, the observed Type I error rate overestimated the nominal significance level. The effect of variance inequality on the statistical power of the rank transform ANOVA when the sampled distribution was symmetric is presented in the next section.

Statistical Power

The proportion of hypotheses rejected when the populations differed by a small, medium, or large effect size (.2, .5, or.8 pooled standard deviation units respectively) are reported in Tables 4 and 5 for the symmetric distributions studied when samples were (20,20) and (17,23) respectively. With equal sample sizes the Brown-Forsythe and parametric ANOVA provided comparable power estimates for all three symmetric distributions. When sample sizes were unequal the Brown-Forsythe procedure provided a more sensitive test for the difference in population means for all three distributions. These results were expected since under the conditions studied with unequal sample sizes, the F-ratio leads to a conservative test.

Differences between power estimates for the rank transform ANOVA and those provided by the Brown-Forsythe and the parametric ANOVA procedures were similar when sample sizes were equal or unequal. For the normal and platykurtic distributions, the rank transform ANOVA provided power estimates slightly lower than those of the other two procedures. When the sampled distribution was leptokurtic, however, the rank transform procedure provided a more sensitive test for the difference in population means than either the Brown-Forsythe or the parametric ANOVA.

Conclusions

The results of the study indicate that the rank transformation approach to analysis of variance can

Table 4

Proportion of Hypotheses Rejected for the Brown-Forsythe (BF), Parametric ANOVA (F) and the Rank Transform ANOVA (RF)

Effect Size	Variance Ratio $\sigma_1^2:\sigma_2^2$	Distribution: Normal BF	Normal F	Normal RF	Platykurtic BF	Platykurtic F	Platykurtic RF	Leptokurtic BF	Leptokurtic F	Leptokurtic RF
Small	1:1.0	.098	.098	.096	.100	.101	.089	.100	.100	.115
	1:1.5	.096	.099	.103	.095	.096	.089	.065	.067	.091
	1:2.0	.095	.099	.096	.084	.086	.078	.099	.100	.106
	1:2.5	.087	.092	.089	.109	.110	.097	.094	.097	.096
	1:3.0	.101	.102	.109	.096	.099	.088	.107	.110	.131
	1:3.5	.091	.092	.099	.102	.104	.095	.095	.097	.116
	1:4.0	.084	.084	.090	.097	.101	.084	.110	.115	.109
Medium	1:1.0	.377	.378	.342	.304	.304	.265	.376	.378	.417
	1:1.5	.341	.342	.319	.328	.329	.287	.348	.351	.386
	1:2.0	.335	.344	.324	.324	325	.275	.364	.364	.410
	1:2.5	.335	.338	.330	.326	.330	.269	.336	.339	.408
	1:3.0	.357	.363	.351	.345	.348	.281	.334	.344	.405
	1:3.5	.318	.326	.312	.298	.302	.236	.370	.373	.443
	1:4.0	.364	.369	.349	.341	.348	.273	.343	.352	.441
Large	1:1.0	.718	.718	.678	.699	.700	.640	.697	.701	.766
	1:1.5	.696	.697	.662	.705	.707	.623	.725	.728	.806
	1:2.0	.678	.682	.644	.689	.694	.611	.719	.724	.798
	1:2.5	.696	.702	.656	.668	.676	.554	.696	.703	.776
	1:3.0	.682	.690	.637	.671	.678	.567	.697	.703	.786
	1:3.5	.683	.689	.661	.655	.663	.537	.706	.721	.800
	1:4.0	.666	.673	.641	.689	.694	.553	.689	.697	.779

Note: $n_1 = n_2 = 20$

Table 5

Proportion of Hypotheses Rejected for the Brown-Forsythe (BF), Parametric ANOVA F, and the Rank Transform ANOVA (RF)

		Distribution								
	Variance Ratio	Normal			Platykurtic			Leptokurtic		
Effect Size	$\sigma_1^2:\sigma_2^2$	BF	F	RF	BF	F	RF	BF	F	RF
Small	1:1.0	.097	.095	.090	.082	.082	.082	.078	.078	.094
	1:1.5	.115	.108	.110	.087	.078	.077	.080	.072	.097
	1:2.0	.104	.089	.083	.093	.076	.086	.098	.082	.120
	1:2.5	.095	.070	.086	.080	.064	.071	.100	.077	.117
	1:3.0	.089	.067	.083	.102	.084	.091	.125	.099	.143
	1:3.5	.102	.074	.088	.104	.079	.089	.119	.091	.112
	1:4.0	.101	.085	.091	.101	.073	.075	.112	.088	.113
Medium	1:1.0	.350	.342	.327	.304	.310	.281	.369	.372	.420
	1:1.5	.331	.314	.304	.356	.329	.296	.382	.367	.429
	1:2.0	.384	.355	.339	.358	.314	.278	.365	.247	.417
	1:2.5	.365	.321	.326	.365	.327	.284	.389	.344	.437
	1:3.0	.377	.324	.348	.350	.299	.252	.389	.337	.448
	1:3.5	.383	.324	.349	.356	.301	.270	.390	.345	.448
	1:4.0	.366	.295	.309	.371	.294	.260	.411	.346	.448
Large	1:1.0	.666	.672	.625	.693	.698	.637	.671	.674	.762
	1:1.5	.707	.680	.662	.691	.669	.617	.731	.718	.792
	1:2.0	.742	.700	.699	.711	.683	.620	.753	.712	.805
	1:2.5	.738	.688	.687	.742	.698	.610	.753	.715	.828
	1:3.0	.754	.679	.689	.737	.685	.610	.772	.714	.819
	1:3.5	.743	.688	.675	.736	.679	.598	.740	.700	.790
	1:4.0	.772	.708	.712	.741	.688	.592	.761	.709	.834

Note: $n_1 = 17$, $n_2 = 23$

provide a solution to the Behrens-Fisher problem, but this solution is appropriate only for a limited set of conditions. In particular the rank transform ANOVA may be recommended when sample frequencies are positively related to group variances and the form of the population distribution is leptokurtic. Under that condition the actual Type I error rate does not overestimate the nominal significance level and the rank transform provides a slight power advantage over the Brown-Forsythe solution. This result was interesting in that the power advantage for the rank transform procedure was obtained even though the actual Type I error rate underestimated slightly the nominal significance level. These results indicate that Type I error rate alone should not be used to evaluate or compare statistical analysis strategies. On the other hand, consideration of both statistical power and actual Type I error rates do provide minimum criteria in judging the usefulness of analysis alternatives.

For other symmetric distributions the rank transform procedure did not provide any statistical advantage compared to the Brown-Forsythe procedure. With skewed population distributions, however, the rank transform approach overestimated the nominal significance level even when the sample frequencies were equal. This finding may be viewed as an important limitation of the rank transform strategy.

As a general solution to the group variance inequality problem, the results of this study do not provide sufficient evidence to recommend any single analysis approach. Before computing hypothesis tests, researchers should first obtain descriptive summary statistics to determine the sample distribution characteristics and to use this information to guide their choice of analysis procedures. For most situations where the population variances differ, the Brown-Forsythe procedure can be used to compare means. This procedure has been shown in the present study, as well as previous investigations, to be generally robust to variance inequality and to provide statistical power comparable to or greater than parametric analysis of variance. There is some evidence however which indicates that when the sampled

distributions are both skewed and leptokurtic, the Brown-Forsythe procedure can overestimate the nominal significance level if sample frequencies and group variances are negatively related.

References

Algina, J., & Olejnik, S. F. (1984). Implementing the Welch-James procedure with factorial designs. Educational and Psychological Measurement, 44, 39-48.

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

Brown, M. B., & Forsythe, A. B. (1974b). The ANOVA and multiple comparisons for data with heterogeneous variances. Biometrics, 30, 719-724.

Clinch, J. J., & Keselman, H. J. (1982). Parametic alternatives to the analysis of variance. Journal of Educational Statistics, 7, 207-214.

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

Conover, W. J., & Iman, R. I. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician, 35, 124-129.

Conover, W. J., & Iman, R. C. (1982). Analysis of covariance using the rank transformation. Biometrics, 38, 715-724.

Dauphin, M. (1983). Level of significance of the rank-transformed t-test in normal populations with unequal variances. Paper presented at the Conference of the American Statistical Associaton.

Dijkstra, J. B., & Werter, P. S. P. J. (1981). Testing the equality of several means when the population variances are unequal. Communication Statistics - Simulation Computation, B10, 557-569.

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

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

Glazer, H. (1963). Comparison of the Student two-sample t-test and the Wilcoxon-Mann-Whitney test for normal distributions with unequal variances. Dissertation Abstracts, 24/05, 2054.

Iman, R. L. (1974). A power study of a rank transform for the two-way classification model when interaction may be present. The Canadian Journal of Statistics. Section C: Applications, 2, 227-239.

Iman, R. L., & Conover, W. J. (1979). The use of the rank transformation in regression. Technometrics, 21, 499-509.

Iman, R. L., Hora, S. C., & Conover, W. J. (1984). Comparison of asymptotically distribution-free procedures for the analysis of complete blocks. Journal of the American Statistical Association, 79, 674-685.

James, G. S. (1951). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. Biométrika, 38, 19-43.

Johansen, S. (1980). The Welch-James approximation of the distribution of the residual sum of squares in weighted linear regression. Biometrika, 67, 85-92.

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

Lee, H., & Fung, K. Y. (1983). Robust procedures for multi-sample location problems with unequal group variances. Journal of Statistical Computation Simulation, 18, 125-143.

Levy, K. (1978). An empirical comparison of the ANOVA F-test with alternatives which are more robust against heterogeneity of variance. Journal of Statistical Simulation, 8, 49-57.

Nath, R., & Duran, B. S. (1981a). A generalization of the rank transform in the two-sample location problem. Communication Statistical-Theory and Methods, A10, 1437-1455.

Nath, R., & Duran, B. S. (1981b). The rank transform in the two-sample location problem. Communication Statistical-Simulation Computation, B10, 383-394.

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.

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

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

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

Wetherill, G. B. (1960). The Wilcoxon test and non-null hypotheses. Journal of the Royal Statistical Society, 23, 402-418.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. Biometrika, 61, 165-170.

AUTHORS

STEPHEN F. OLEJNIK, Associate Professor, Department of Educational Psychology, University of Georgia, Athens, Georgia 30602

JAMES ALGINA, Professor, Foundations of Education, University of Florida, Gainesville, Florida 32611