TWO EXTENSIONS OF LEVENE'S TEST OF HOMOGENEITY OF VARIANCES: A MONTE CARLO INVESTIGATION OF ACTUAL-NOMINAL ERROR RATE AND RELATIVE POWER

Dissertation for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
JAMES DOUGLAS MAAS
1975

This is to certify that the

thesis entitled

TWO ESTENSIONS OF LEVENE'S TEST OF HOMOGENEITY
OF VARIANCES: A MONTE CARLO INVESTIGATION
OF ACTUAL-NOMINAL ERROR RATE AND
RELATIVE POWER

presented by

James Douglas Maas

has been accepted towards fulfillment
of the requirements for

PH.D. degree in SECONDARY ED. & CURRICULUM

Maryellen McSweeney
Major professor

Date 7 February 1975

ABSTRACT

TWO EXTENSIONS OF LEVENE'S TEST OF HOMOGENEITY OF VARIANCES: A MONTE CARLO INVESTIGATION OF ACTUAL-NOMINAL ERROR RATE AND RELATIVE POWER

By

James Douglas Maas

This study was motivated by the need for a good K-sample test for equal variability. Well over a dozen tests have been suggested in the literature but all suffer from one or more of the following ailments:

(1) Poor correspondence between nominal type I error rate and empirical type I error rate if the normality assumption is violated;

(2) Low power when this correspondence is good;

(3) Lack of post hoc procedures associated with the method.

Of all the tests which have been considered in the recent literature, only two have shown an acceptable correspondence between the nominal and empirical type I error rates, no matter what distribution is sampled from. These are Box's test and Moses' test. Box's test was

found to be more powerful than Moses' test for all cases considered. Unfortunately, neither of these procedures is desirable for sample sizes much less than fifteen.

Of the remaining tests, one suggested by Levene showed the most promise. Levene's test consists of an analysis of variance on the absolute deviations of the observations from their sample means. The average amount by which the empirical and nominal type I error rates differed was relatively low for this test; however, Levene's test consistently gave somewhat liberal estimates of the nominal alpha. Both the Kruskal-Wallis test and the Normal Scores test of location are known to be somewhat more conservative than is ANOVA for most distributions sampled. Usually little power is sacrificed when performing the Kruskal-Wallis test or the Normal Scores test. It was thought that the conservative nature of these tests might counter the liberal nature of Levene's test to produce a test statistic which was acceptable in its empirical type I error rate. Both the Kruskal-Wallis extension and the Normal Scores extension of Levene's test would hopefully have power comparable to that of Levene's test. Thus, the thrust of this study was a Monte Carlo investigation of the properties of both the Kruskal-Wallis extension and the Normal Scores extension of Levene's test for equal variability.

Two, three and four levels of the independent variable were considered for various small sample sizes. The properties of the test statistics were observed when sampling from normal, uniform and exponential distributions. For each of the cases a simulated analysis was repeated one thousand times and the number of rejections of the null hypothesis of equal variability was counted using a series of commonly employed nominal alpha levels. Both nominal-empirical alpha level fit and power were of concern in this study.

When samples were taken from either uniform or normal distributions, the results were in the direction predicted. Both extensions of Levene's test produced better estimates of type I error rate than did Levene's test. The Normal Scores extension proved to give generally better estimates of type I error rate and power than did the Kruskal-Wallis extension. For $\alpha = .10$ and $\alpha = .05$, both extensions were still somewhat liberal and for $\alpha = .01$ they were slightly conservative. In these same situations, Levene's test had slightly more power; however, some of the advantage with respect to power may be attributable to the slightly greater liberality of Levene's test.

When the exponential distribution was sampled from, the empirical estimates of α were all very liberal. With this poor quality fit, the power comparisons were

meaningless. With the failure of these techniques for the exponential distribution, the good K-sample test for equal variability which is somewhat distribution free had not been found. This failure indirectly suggested the use of a similar but slightly different technique.

The properties of a modified form of Levene's test and of the two extensions were observed in a mini-study. The test statistics were computed as before, with the exception of deviating observations from sample medians rather than sample means. The median deviation technique was employed for the three sample design with six observations per sample when sampling from normal distributions and when sampling from exponential distributions. The results of this mini-study were quite encouraging.

For $\alpha = .10$ and $\alpha = .05$, the empirical type I error rates of the Levene type test were quite close to the nominal level when sampling from either distribution. However, for $\alpha = .01$ the empirical type I error rates were too liberal. With the exception of Box's test and Moses' test, this is the only time that the type I error rate was well behaved when sampling from either a distribution with high kurtosis or with heavy skew. Although the Levene type test using the median was not without power, it appears to be considerably less powerful than the appropriate conventional tests when sampling from

normal distributions. The Kruskal-Wallis median extension and the Normal Scores median extension of the Levene type test did not fare nearly as well in either their empirical estimates of type I error rate or power.

TWO EXTENSIONS OF LEVENE'S TEST OF HOMOGENEITY OF VARIANCES: A MONTE CARLO INVESTIGATION OF ACTUAL-NOMINAL ERROR RATE AND RELATIVE POWER

By

James Douglas Maas

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Secondary Education and Curriculum

1975

ACKNOWLEDGMENTS

One seldom gets an opportunity to publicly say thanks to the people who have profoundly and positively affected his life. I would like to take this space to do just that.

To Dr. Maryellen McSweeney, who directed this dissertation effort, thank you for the faith that you showed in me. It has been a pleasure to work closely with a tireless, brilliant and dedicated teacher.

To Dr. Ted Ward, who chaired my doctoral committee, my thanks for being so understanding. My co-teaching experience with you will remain among my most pleasant memories.

To Dr. Howard Teitelbaum, Dr. Marvin Grandstaff and Dr. Andrew Porter, my thanks for your constant guidance and encouragement as committee members.

A doctoral program would be very difficult without good teachers and good friends. All of these people I hold in the highest regard not only as intellectual leaders but also as friends.

Finally, to my wife Jean, a super thanks to a girl who truly earned her Ph.T. Degree. Thanks for keeping me on the right track to the very end. Hopefully, the future will be bright for you because of all the hardships you endured the past four years.

TABLE OF CONTENTS

LIST OF TABLES

Ta

5

5-

5-

5-

5-1

5-1

5-1

5-11

5-14

Table Page

Table | Page

c

i

b

b

v

q

h.

tc

na

me

co

er

CHAPTER I

RATIONALE

An Observation

In educational research, there are often analysis problems which are a function of the fact that the real world is not neatly structured to meet the assumptions required in the models that researchers have available to them. When faced with a situation which knowingly does not conform to the models available, one can react in different ways. The most common approach seems to be to use the model knowing the assumptions are probably being violated. If the test is robust with respect to violation of these assumptions, then this approach is quite sound. In many of the common situations robustness has been investigated and this information is available to the researcher.

Should the test be nonrobust there are two alternatives one has if he wishes to make inferential statements. The model and its associated test of significance could be used without knowledge of how close the actual error rates are to the nominal error rates when the w

t

t

c

s

di

in

(2

te

of

sit

rec

Grc

particular violation is made. For many researchers this may be the only route available.

If the violation of assumptions seems critical, another option consists of the development of a new model which does not demand these assumptions. This may be as "simple" as a data transformation or as complex as the development of a new test statistic and its distributional properties. Often this is a difficult task with many dead ends. Occasionally the model is effective and all ends on a happy note.

An Example

Recently, an analysis situation was encountered which did not conform to a known model (26). A researcher, teaching prospective counselors to use behavioral objectives effectively in the counseling situation, wished to compare the effectiveness of three techniques. Thirty-six subjects were randomly assigned to one of three conditions. Group A received (1) an article in which the importance of behavioral objectives was emphasized, (2) a manual which explained conditions, criterion and terminal behavior statements and which gave illustrations of each and (3) a rating form to be used in comparing situations as to their use of these objectives. Group B received the article and manual but not the rating form. Group C received only the article.

After getting acquainted with the above materials, each subject was placed in a situation in which he was asked to compare each of three pairs of audiotapes on their use of behavioral objectives. For each pair he was to pick the better of the two. Group A was able to use the rating form to assist them whereas Groups B and C were asked to make their choice without the use of this rating form. In fact, at no point in the experiment did Groups B and C have access to the rating form. In each group half of the subjects received feedback on their pair selections and half did not. Again, those to get feedback were selected at random.

After this experience each subject was placed in a simulated counseling situation. The dependent variable of interest was the number of uses of behavioral objective statements per unit time that was observed during the simulated situation. The design was:

	Group A	Group B	Group C
	T_1	T_2	T_3
Feedback F_1	XXX XXX	XXX XXX	XXX XXX
No Feedback F_2	XXX XXX	XXX XXX	XXX XXX

Thirty-six subjects were randomly assigned to the six cells so that there were six observations per cell. Two of the research questions of interest were:

1. Does Treatment A produce a more homogeneous group than Treatments B and/or C?

2. Are those groups receiving feedback more homogeneous than those groups not receiving it?

Analysis Techniques Available

A search was made for a good test which was created specifically to answer questions about homogeneity of variances for designs with two or more design variables. The following three criteria were used to judge the quality of prospective multi-factor tests of equal variability. First, there must be good agreement between the nominal α and actual α. This means that the probability of rejecting the null hypothesis is actually α when the test is performed at the nominal α level. Second, the test must possess good power. This means that there is a high probability of rejecting the null hypothesis when it is not true. Third, the test should have post hoc procedures associated with it. This means that if the null hypothesis is rejected, techniques are available to locate the sources of rejection and to estimate their magnitude.

The search for a multifactor test of variance homogeneity meeting these criteria was fruitless. Those multifactor tests which were identified suffered from serious limitations. The tests developed by Russell and Bradley (53), Han (23) and Shukla (53) are restricted to multifactor designs with no interaction effects and with a single source of variance heterogeneity. The test proposed by Overall and Woodward (49) is less restrictive in its design specifications, but is believed to be sensitive to population nonnormality. Simultaneous post hoc confidence interval procedures are not available for any of these tests except that of Overall and Woodward, and the small sample properties of the tests have not been investigated for nonnormal data. Although the literature on multifactor tests for variance homogeneity was limited and the tests were disappointing, the literature on methods of comparing variances for single factor designs was somewhat more extensive.

The above design could be considered as a single factor design if the six cells were treated as six levels of one design variable as pictured below:

T_1 F_1	T_1 F_2	T_2 F_1	T_2 F_2	T_3 F_1	T_3 F_2
XXX XXX	XXX XXX	XXX XXX	XXX XXX	XXX XXX	XXX XXX

r

s

a

c

m

e

bu

hi

Bo

Bartlett, Cochran, Hartley, Cadwell, Kendall, Scheffé, Bargmann, Box, Levene and others have addressed themselves to the development of analysis techniques for testing several populations for equality of variance. However, all of these methods suffer from one or more of the following ailments:

(1) Poor correspondence between nominal type I error rate and actual type I error rate if the normality assumption is violated;

(2) Low power when this correspondence is good;

(3) Lack of post hoc procedures associated with the method.

In a 1972 article, Gartside (15) compared six different methods of testing for homogeneity of variance. These six methods are (1) Bartlett's statistic, (2) a modification of Bartlett's statistic, (3) Cochran's statistic, (4) Cadwell's statistic, (5) Box's statistic and (6) Bargmann's modification of Box's statistic.

Under the normality and equal size sampling cases he found that all six methods very favorably matched nominal type I error rate and actual type I error rate. However, under the condition of nonnormality but equal sample size all but Box's test were markedly high in their actual type I error rates. Unfortunately Box's test proved to have less power under these same conditions. For one alternative ($\sigma_1^2 = C° = 1, \sigma_2^2 = C^1 = C, \ldots \sigma_K^2 = C^{K-1}$) Bartlett's statistic was more powerful and for another ($\sigma_1^2 = \sigma_2^2 = \ldots \sigma_{K-1}^2 = 1, \sigma_K^2 = C$) Cochran's was more powerful. As a result of this article one could conclude that the traditional tests left much to be desired.

In 1960 Levene devised a test based on ANOVA of absolute deviations from the mean (36). Using Monte Carlo techniques, Levene showed, for equal sized samples of ten or twenty per cell with two and three cell designs, that the empirical type I error rate and nominal type I error rate practically coincided. This was true for normal, uniform and double exponential distributions. At the nominal .05 level, the largest empirical probability obtained was .063.

Levene also investigated the power of his test for rejecting H_o and found it to be quite satisfactory. Assuming normal distributions, the approximate efficiency of Levene's test with respect to Bartlett's test ranged from .83 to .90 depending on the value of alpha and the alternative hypothesis considered. In addition to these properties, Levene's test, employing a fixed effects ANOVA model, has techniques available to locate specific differences in variability should H_o be rejected.

This test went relatively unnoticed by applied statisticians until a 1966 American Educational Research

Journal (17) article by Glass brought it to the forefront. From 1972 to 1974, three articles were written indicating that Levene's test has inflated type I empirical probabilities for cases in which the cell sizes are small or the number of cells is large. This indicates a need for more exploration into possible K sample tests which have (a) correspondence of nominal and empirical type I error rate better than that of Levene's test, (b) available post hoc procedures, (c) more power than Levene's test and (d) efficiency close to Bartlett's test in the situation for which Bartlett's test was designed.

Two Proposed Tests

In multicell tests of location, ANOVA is traditionally the test used. It has been found to be robust with respect to violation of the normality assumption and, if cell sizes are equal, the homoscedasticity assumption (18). However, ANOVA has some genuine competitors. The Kruskal-Wallis (33) test is one which is rather insensitive to differences in the skewness, kurtosis or scale of the K population distributions. This test is insensitive to nonlocation alternatives and has a high asymptotic efficiency relative to the ANOVA F test as a test for location alternatives (.955 when the assumptions of normality and common variance are satisfied). If the underlying distributions are nonnormal the efficiency relative to the ANOVA F test often equals or exceeds one. Two such nonnormal distributions for which the Kruskal-Wallis test is at least as efficient as the ANOVA F test are the uniform and the exponential distribution.

The normal scores test is yet another good competitor with ANOVA (39). When the assumptions of normality and common variance are met, the asymptotic efficiency of the normal scores test relative to the ANOVA F test is one. It is more efficient than the Kruskal-Wallis test when the samples have been drawn from the uniform or exponential distributions (27). Both the Kruskal-Wallis test and the normal scores test have been shown to be somewhat conservative in the empirical estimates of type I error rate.

It has been stated that (1) Levene's test appears to be an efficient test relative to Bartlett's test, (2) Levene's test is somewhat liberal in its empirical estimates of actual type I error rate, but (3) the Kruskal-Wallis test and normal scores test are very good competitors to ANOVA in tests of location and are even more efficient than ANOVA if the normality and homoscedasticity assumptions are not met, and (4) the Kruskal-Wallis test and the normal scores test are somewhat conservative in their empirical estimates of actual type I error rate. This would suggest two other possible methods of testing K populations for differences in variance. The first is a Kruskal-Wallis extension of Levene's test and the second is a normal scores extension of Levene's test.

First, consider Levene's test. What Levene suggests is to calculate Z_{ij}, the absolute value of the difference of each observation in a given cell from the cell mean. For observation i in cell j, $Z_{ij} = |X_{ij} - \overline{X}_{.j}|$. Then perform a one-way ANOVA on the Z_{ij} scores. The test statistic is the ratio of $MS_{between}$ to MS_{within} of the transformed Z_{ij} scores. In a J cell design with a grand total of N observations, if this F-ratio exceeds the (1-α) percentile point of the central F distribution which has J-1 and N-J degrees of freedom, one concludes, at the α significance level, that the populations are not identical in variability.

The first extension suggested is based on the Kruskal-Wallis statistic. This extension involves the following steps:

1. Calculate $Z_{ij} = |X_{ij} - \overline{X}\ ._{j}|$ for each individual.
2. Rank order these Z_{ij} scores from 1 to N and replace each Z_{ij} with its corresponding rank r_{ij}.
3. Obtain the mean rank $\overline{r}_{j}$ for each cell.

4. Compute the Kruskal-Wallis statistic.

$$H = \frac{12}{N\ (N+1)} \sum_{j=1}^{K} n_j \bar{r}_j^{\,2} - 3(N+1)$$

5. If the H statistic exceeds the (1-α) percentile point of the central χ^2 distribution with K-1 degrees of freedom, conclude, at the α significance level, that the populations are not identical in variability.

The second extension that is suggested is a K-sample normal scores extension of Levene's test. The following steps are used in this extension:

1. Calculate Z_{ij} scores for each individual.

2. Rank order the Z_{ij} scores from 1 to N and replace each Z_{ij} with its corresponding rank r_{ij}.

3. Replace each r_{ij} with its inverse normal score $\phi^{-1}_{ij} = \phi^{-1}(r_{ij}/(N+1)\ )$ where $\phi^{-1}(p)$ denotes the point on the standard normal distribution which cuts off the cumulative probability p.

4. Calculate the test statistic W where

$$W = \frac{(N-1)\ \sum_{j} \left(\left(\sum_{i} \phi_{ij} \right)^2 / \ n_j \right)}{\sum\sum_{ij} \ (\phi^{-1}_{ij})^2}$$

5. If W exceeds the (1-α) percentile point of the central χ^2 distribution with K-1 degrees of freedom, conclude, at the α significance level, that the populations are not identical in variability.

The Research Questions

The exact questions to be answered by this study are: For the Kruskal-Wallis extension of Levene's test and the Normal Scores extension of Levene's test, (1) How do the nominal type I error rate and the actual type I error rate correspond? (2) How powerful are these tests with respect to Levene's test? and (3) For the two-sample case how powerful are these tests with respect to the traditional parametric test $F = s_1^2/s_2^2$?

Two methods, the analytic and the empirical, can be used to obtain the sampling distribution of test statistics such as F used in Levene's case or H or W in the two extensions suggested. When possible the analytic method is preferred because the distributions involved are exact and free from sampling error. Levene was forced to use empirical Monte Carlo techniques "because of the well known difficulty of obtaining the distribution of F from a non-normal population analytically." The difficulty involved absolute value integration and further complications which he mentions (36, p. 280).

For the same reasons, the questions of this study must be answered by an empirical sampling (Monte Carlo) method rather than analytically.

An Overview

The current tests of scale, including Levene's test are reviewed in Chapter II. In addition, the properties of ANOVA, the Kruskal-Wallis test and the Normal Scores test are compared. The primary questions of the study and the techniques used to answer them are found in Chapter III. These techniques involve the generation of uniform, normal and exponential random variates, inverse normal scores and some F values. The generators used are discussed in Chapter IV. The results of the study are found in Chapter V. These results coupled with a finding of Miller (42) suggested that a modification of Levene's technique might prove promising. A mini-study exploring this modification and the results of the mini-study are also found in Chapter V. The study is summarized and the primary conclusions are listed in Chapter VI. A discussion follows in which other possible techniques for testing equal variability are explored.

CHAPTER II

REVIEW OF THE LITERATURE

K Sample Tests of Scale ($K>2$)

Since 1931, statisticians have been searching for a good test for homogeneity of variability for K populations. One important criterion for a good test is that it be relatively distribution free. This means that no matter what distributions are sampled, the proportion of rejections, when the null hypothesis of equal variance is true, is close to the nominal level of significance of the test. If more than one test should have this property, then the one with the greatest power is typically preferred.

In this section, tests proposed by Bartlett (1), Hartley (24), Cochran (31), Cadwell (7), Box (3), Box and Andersen (4), Moses (44), Burr (12), Miller (42) and Levene (36) are reviewed. The small sample properties of these tests have been examined by means of Monte Carlo techniques by Games et al. (14), Gartside (15), Hall (21), Layard (34), Levene (36), Miller (42) and Winslow and Arnold (63). A summary of the tests each studied the distributions sampled, the designs used, and the power alternatives used is displayed in Table 2-1.

Gartside (15) claimed that the first approach to the K sample test of equal variability was made in 1931 by J. Neyman and E. Pearson using the likelihood ratio statistic approximately distributed as χ^2_{K-1}. Bartlett (1) modified this statistic to improve the approximation to the chi-square distribution. His test statistic is:

$$\beta = \frac{(N-K)\ \ln\ (\Sigma v_i s_i^2/\ (N-K))\ -\ \Sigma v_i \cdot \ln\ s_i^2}{1 + \frac{1}{3(K-1)}\ (\Sigma \frac{1}{v_i} - \frac{1}{\Sigma v_i})}$$

where:

N = total sample size

K = number of samples

s_i^2 = variance of sample i and

v_i = degrees of freedom upon which s_i^2 is based.

This statistic is referred to the chi-square distribution with K-1 degrees of freedom. Needless to say, the computations involved with this test are quite tedious. Box (3) first noticed that if nonnormal populations are sampled when using Bartlett's test, the probability of a type I error may be much larger or smaller than the nominal α. The direction of the departure from the nominal significance level depends on the kurtosis of the populations sampled.

TABLE 2-1.--Monte Carlo studies of tests of scale in recent literature

TABLE 2-1.--Monte Carlo studies of tests of scale in recent literature

Authors	Tests Considered	Distributions Sampled From	Designs Used		Power Alternatives Used
			# Samples	Sample Size	
			2	20	(1) $\sigma_1=1$ (2)
Levene	Levene z	Normal	2	10	$\sigma_2=1.3$ $\sigma_1 \ . \ . \ . \ \sigma_4=1$
1960		Uniform	4	20	$\sigma_3=1.6$ $\sigma_5 \ . \ . \ . \ \sigma_7=1.3$
	Levene z^2	Dbl Expon	10	20	$\sigma_4=2.0$ $\sigma_8 \ . \ . \ .\sigma_{10}=1.6$
					$\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2$
Layard	Bartlett	Uniform			
1968	Box	Normal	4	10	(1) 1, 1, 2, 2
	Jackknife	Dbl Expon	4	25	(2) 1, 2, 3, 4
					(3) 1, 1, 4, 4
	Box	Normal			
Miller	Box-Andersen	Uniform	2	10	
1968	Moses	Dbl Expon			$\sigma_2 = \sigma_1 + C$
	Jackknife	Sk Dbl Expon	2	25	
	K=1, K=5	Sixth Power			
	Bartlett				
	Box	Normal			
Hall	Box-Andersen	Uniform	5	10	
1972	Moses	Dbl Expon			$\sigma_K = c+(K-1)d$
	Jackknife K=1&5	Sk Dbl Expon	5	25	
	Levene z^2	Sixth Power			

TABLE 2-1.--Continued

Authors	Tests Considered	Distributions Sampled From	Designs Used		Power Alternatives Used
			# Samples	Sample Size	
			3	4	
	Bartlett		4	4	
Gartside	Box		5	4	(1) $\sigma_K^2 = c^{K-1}$
1972	Cadwell	Normal	10	4	
	Cochran	Weibull	3	16	(2) $\sigma_1^2 = \sigma_2^2 = . . . \sigma_{K-1}^2=1$
	Hartley		4	16	
			5	16	$\sigma_K^2=c$
			10	16	
	Bartlett				
Games	Box				
Winkler	Box-Andersen	Normal			
Probert	Burr	Uniform	3	6	
1972	Cochran	Slight Skew			$\sigma_K = c + (K-1)d$
	Hartley	Moderate Skew	3	18	
	Levene z_2	Extreme Skew			
	Levene z^2	Sym Leptokurtic			
Winslow +			2	6	
Arnold	Moses	Normal	2	12	
1972			2	18	$\sigma_2 = \sigma_1 + C$
			2	24	

This was further confirmed by Layard (34), Hall (22), Gartside (15) and Games et al. (14). Layard looked at four sample designs with equal cell sizes of ten and twenty-five. He sampled from normal, uniform and double exponential distributions. Hall sampled from these same three distributions and looked at five sample designs with equal cell sizes of ten and twenty-five.

In addition he sampled from a skew double exponential distribution and a sixth power distribution. Gartside sampled from the normal and the Weibull distributions for a variety of designs with equal cell sizes of four and sixteen. Games et al. looked at three sample designs with equal cell sizes of six and eighteen and samples from normal, uniform, slightly skewed, moderately skewed, extremely skewed, and symmetric leptokurtic distributions.

Hartley (24) developed the statistic F_{max} = (largest of K variances) / (smallest of K variances). The F_{max} distribution is tabled in Kirk's text (31). Box (3) showed that as with Bartlett's test, this test will be liberal when sampling from leptokurtic distributions and conservative when sampling from platykurtic distributions. The Monte Carlo studies of Games et al. (14) and Gartside (15) support this contention.

Cochran (31) devised a relatively simple test statistic, the ratio of the largest of the K cell variances to the sum of all the cell variances. Critical values of Cochran's statistic are tabled in Kirk (31). Unfortunately, Box (3) showed once again that with non-normal distributions, this test has the same problems as Bartlett's test and Hartley's test. The Monte Carlo studies of Gartside (15) and Games et al. (14) support the claim of Box.

Cadwell (7) devised a statistic which is the ratio of the largest of the sample ranges to the smallest of the sample ranges. It is referred to the χ^2 distribution. Gartside (15) found this test to be too liberal when sampling from the Weibull distribution. This was the only distribution he considered other than the normal distribution.

Box and Andersen (4) used permutation theory to modify Bartlett's test. They define the statistic:

$$M^1 = \frac{B}{(1 + G/2)}$$

where:

B = Bartlett's statistic and

G = a function of the sample estimate of the kurtosis.

They investigated the behavior of their test and found that for normal, uniform and double exponential distributions it gave empirical probabilities closer to the nominal α than Bartlett's test when the equal variance hypothesis was true. This study was based upon only two hundred samples. Using a three-cell design with six observations per cell, Games et al. (14) found the Box-Andersen test was too liberal for four out of the six distributions considered in their study. Surprisingly, they found the test to be very conservative when sampling from the uniform distribution whereas Box and Andersen had found their test to be somewhat liberal for this case. Hall (22) also found the Box-Andersen statistic very liberal for all five distributions he used when sampling ten observations per sample from a five-sample design. The test was found to behave better with twenty-five observations per sample; however, the empirical estimates of type I error rates ranged from .025 to .087 depending upon the distribution sampled. Miller's (42) results were similar to those of Hall. The only difference between the studies of Miller and Hall was the number of samples in the design; Miller chose two samples, and Hall chose five samples.

Bartlett and Kendall (2) suggested a test using ln s^2, and Scheffé (51) described this test in his classic text. Box (3) further explored the nature of this test, which has been cited in the recent literature as the Box test. The test consists of breaking each of the K samples into subsamples, computing ln s^2 for each subsample, and doing an ANOVA on the variable ln s^2 to test equality of the population variances. Bartlett and Kendall warned that the quality of the test may be poor for small cell sizes. Miller (42) looked at the quality of this test for two sample designs with twenty-five observations per sample. Each sample was subdivided into five subsamples each of size five. For the distributions Miller sampled, the test appeared to be well behaved at $\alpha = .05$ and $\alpha = .01$. The largest empirical discrepancy from the nominal α was .009 for the skew double exponential distribution at nominal $\alpha = .05$.

Games et al. (14) looked at the quality of this test for three sample designs with eighteen observations per sample and two different subsampling patterns. The first sampling pattern consisted of subdividing the eighteen observations within a sample into nine subsamples of two each; the second used six subsamples of three each. They considered six different distributions with varying degrees of skew and kurtosis. Both subsampling patterns yielded consistently conservative estimated probabilities of type I errors. The power estimates were considerably higher for the second variation than for the first. Games et al. concluded that the second variation had reasonable power for all populations sampled from at $\alpha = .05$.

Moses (44) proposed a nonparametric modification of the Box test. It is simply the Kruskal-Wallis test applied to the value of ln s^2 obtained from subsamples as in the Box test.

For the two-sample situation with twenty-five observations per cell, Miller (42) found that with subsamples of size five in each sample, the Moses test was somewhat conservative in its estimates of type I error rates and was consistently less powerful than Box's test. For the five-sample design with twenty-five observations per sample, Hall (21) obtained results similar to those of Miller. Winslow and Arnold (63) used two-sample designs with equal sample sizes of 6, 12, 18 and 24 with different subset sizes. They sampled from the normal distribution and obtained somewhat less conservative results.

Miller (42) proposed a jackknife technique to test variance equality in the two-sample design, and Layard (34) extended this test to the K sample design. This test involves computing a one-way ANOVA on the U_{ij} where:

$$U_{ij} = n_i \ln s_i^2 - (n_i-1) \ln s_i^2 (j)$$

where:

n_i = the number of observations in cell i and

s_i^2 = the sample variance in cell i and

$$s^2_{i(j)} = \frac{\sum\limits_{K \neq j} (x_{iK} - \bar{x}_{i(j)})2}{n_i - 2}$$

where:

$$\bar{x}_{i(j)} = \frac{\sum\limits_{K \neq j} x_{iK}}{n_i - 1}$$

Miller found that for two sample designs with twenty-five observations per sample, the empirical estimates of $\alpha = .05$ ranged from .029 to .082 depending on the distribution sampled; with ten observations per cell the estimates ranged from .020 to .126. The power was consistently higher than that obtained using Box's test. In a nearly identical study using five samples, Hall obtained almost identical results. Layard's (34) results were similar when four samples were observed.

Burr and Foster (12) introduced the statistic:

$$Q = \Sigma s_K^4 / (\Sigma s_K^2)^2$$

where:

s_K = the standard deviation in sample K.

Games et al. (14) found it to be a very unstable estimator of type I error rates. For $\alpha = .05$, the empirical estimates ranged from .001 to .371, depending on the distribution sampled.

The average empirical estimates of type I error rates for twelve different K sample tests based on the Monte Carlo studies cited in Table 2-1 are displayed in Table 2-2 for a series of distributions with differing skewness and kurtosis. It can be observed that the only statistics which have reasonable estimates of type I error rates for all the distributions considered are Box's test and Moses' test. In the studies of Hall (22) and Miller (42), Box's test was always a better estimator of type I error rate and was always more powerful than was Moses' test.

In addition to the previously described tests is Levene's test (36) which is a focal point of this study.

Levene's Test of Scale

In 1960, Levene (36) suggested a new technique for testing for equal variability in K populations. His test consists of an ANOVA on the absolute deviations of the observations from their sample means. One aspect of his study was a Monte Carlo investigation of the empirical probability of obtaining a significant F-ratio under the null hypothesis: $\sigma_1 = \sigma_2 = . . . = \sigma_K$. He restricted his study to two-, four- and ten-sample designs with equal sample size of ten or twenty. The distributions sampled from were the normal, the uniform, and the double exponential.

TABLE 2-2.--Average empirical estimates of type I error rates* when α = .05 for twelve K-sample tests of scale based on Monte Carlo studies found in the literature (see Table 2-1)

Test Distribution Skew Kurtosis	Bartlett	Box	Box + Andersen	Burr	Cadwell	Cochran	Hartley	Moses	Jackknife K=1	Jackknife K=5	Levene z	Levene z^2
Uniform 0 -1.2	.006 6	.047 6	.084 5	.001 1	-- --	.009 1	.014 1	.038 2	.028 4	.046 2	.063 4	.054 6
Normal 0 0	.052 12	.049 18	.078 5	.035 1	.049 4	.050 5	.047 5	.049 12	.049 4	.067 2	.057 6	.051 8
Double Expon. 0 3	.246 6	.053 4	.069 4	-- --	-- --	-- --	-- --	.030 2	.070 4	.075 2	.073 3	.046 5
Sk Dbl Expon. ? 6	.233 4	.049 2	.081 4	-- --	-- --	-- --	-- --	.031 2	.075 4	.086 2	-- --	.052 2
Sym Leptokurtic 0 6.16	.278 2	.040 2	.047 1	.334 1	-- --	.197 1	.177 1	-- --	-- --	-- --	.076 1	.048 1
Sixth Power 0 33	.420 4	.048 2	.051 4	-- --	-- --	-- --	-- --	.029 2	.104 4	.108 2	-- --	.036 2
Slight Skew .45 .14	.049 2	.042 2	.083 1	.042 1	-- --	.054 1	.044 1	-- --	-- --	-- --	.088 1	.055 1
Moderate Skew .64 .53	.067 2	.041 2	.082 1	.063 1	-- --	.065 1	.074 1	-- --	-- --	-- --	.082 1	.068 1
Extreme Skew 2.04 6.54	.325 2	.040 2	.076 1	.371 1	-- --	.255 1	.226 1	-- --	-- --	-- --	.143 1	.077 1
Weibull 0 2.71-6.24	.202 8	.052 8	-- --	-- --	.160 8	.192 8	.180 8	-- --	-- --	-- --	-- --	-- --

*Lower number is the number of empirical estimates averaged for that cell of the table.

Levene pointed out the desirability of obtaining the distribution of his test statistic analytically; however, because of the mathematical difficulty of working with nonnormal populations and other problems which he cited, he was forced to use empirical sampling methods. The results of his investigation were very encouraging. The empirical probabilities obtained were for the most part very close to the nominal alpha used (.10, .05 or .01). The only situation for which the fit was quite poor was a ten-sample design with twenty observations per sample when sampling from the double exponential distribution. The empirical probability for $\alpha = .10$ was .154, for .05 it was .085 and for .01 it was .028.

Levene also investigated the power of his test when sampling from normal distributions. For two-sample designs, he found the efficiencies relative to the F test ranged from .75 to .96 for various values of α and various alternative hypotheses. For the K sample case, the efficiency of his test relative to Bartlett's test ranged from .83 to .90.

It was the quality of these results which prompted Glass to write an article (17) in which he commended the test to the attention of the consumers of statistics. Later, in an educational statistics text, Glass and Stanley (19) recommended the test, although they pointed out that the robustness of the technique was demonstrated only for designs with equal cell size.

This test did not come under closer scrutiny until 1972. In a letter to the editor of Technometrics (41), Miller criticized the use of Levene's test in a study which appeared in an earlier issue of that journal. The authors had used the test with smaller sample sizes than Levene had used in demonstrating the robustness of his test. In Levene's Monte Carlo study, he demonstrated that for equal sample sizes of three or four, that the ratios of between group variance to within group variance gave inflated F values. He attributed this inflation to the dependence of the deviations within a sample which is accentuated in small sample sizes.

Games et al. (14) in another Monte Carlo study using a three-sample design with six observations per sample found the Levene test to produce liberal estimates of type I error rates, ranging from .069 to .143 for a nominal α of .05. The other statistics studied had even wider ranges of estimated α, but Levene's test was less powerful than two of the tests with which it was compared.

Neel and Stalling (47) tested the type I error rate for Levene's test for designs with sample sizes ranging from three to twelve and with the number of samples ranging from two to seven. They used ten thousand replications sampled from normal distributions in each of the sixty simulations and the results were indeed disappointing. Every single estimate of $\alpha = .05$ was liberal. "Two interacting trends seem discernable in this upward bias: (a) for smaller n (roughly, n less than 9), the bias increases as the number of variances tested increases; (b) for larger n, the bias seems to remain constant or even to fall as the number of variances increases" (47, p. 5). They concluded with a statement indicating that Levene's test has little to offer the educational researcher and discouraged its use. Their results are partially displayed in Table 2-3.

TABLE 2-3.--Observed number of type I errors in ten thousand tests, given a nominal level of .05

Number of Variances	Sample Sizes				
	4	6	8	10	12
2	791	659	601	576	545
4	959	781	597	583	516
6	1174	902	638	613	663

In exploring the asymptotic properties of Levene's test, Miller (42) concluded that if deviations were taken from the sample means, Levene's test statistic would not be asymptotically distribution free. However, he was able to prove that it would be asymptotically distribution free if deviations were taken from sample medians. He did not explore the small sample properties of this modification in his study because he did not discover this property until all the other work in that study was completed.*

It should be pointed out that in addition to ANOVA of absolute deviations (z_{ij}), Levene looked into the possibilities of ANOVA of z^2, ANOVA of $\sqrt{z}$ and ANOVA of ln z. He concluded that the tests based on z and z^2 are satisfactory while those based on $\sqrt{z}$ and ln z have poor power with even more liberal estimates of type I error rates. His ANOVA based on z^2 has been subjected to further Monte Carlo study by Hall (22) and Games et al. (14). Hall (22) compared Levene's z^2 test with six other tests for equal variability for the five sample design with equal sample sizes of ten and twenty-five. He sampled from five distributions which had kurtosis ranging from -1.2 to 33. For all five distributions, Levene's z^2 test was somewhat conservative for the sample size of twenty-five. The nominal .05 estimates ranged from .023 to .048. However, with a sample size of ten, the nominal .05 estimates ranged from .049 to .067 which are somewhat liberal estimates. For

*A recent Monte Carlo study by Brown and Forsythe has examined this modification. The results of their study are reported in Chapter VI.

distributions with low kurtosis, Levene's z^2 test was relatively powerful but for those distributions which had high kurtosis, it was relatively weak.

Games et al. (14) found Levene's z^2 test to give better estimates of type I error rate than Levene's z test did. For all six distributions studied, Levene's z^2 statistic was less liberal than Levene's z statistic but was considerably less powerful than Levene's z statistic.

K Sample Tests of Location (K>2)

When Levene's statistic is calculated, an ANOVA is performed on the absolute deviations (z_{ij}) from the sample means.

$$z_{ij} = |x_{ij} - \bar{x}_{ij}|$$

The assumptions necessary to use the F distribution with the ANOVA model are:

1. The observations are drawn from normally distributed populations.
2. The observations represent random samples from the populations.
3. The variances of the populations are equal.

If all samples are drawn randomly from identical populations, then assumptions (2) and (3) are met for z_{ij} scores; however, the first assumption is not likely to be met. The z_{ij} scores are unlikely to be normally distributed. Their distribution will be a function of the distribution of the x_{ij} scores.

What then will be the effect of the violation of the first assumption? Glass et al. (18) thoroughly review the consequences of the failure to meet this assumption. In summary, they state that skewed populations have very little effect on either the level of significance or the power of the fixed-effects ANOVA F-test. With respect to kurtosis, they indicate that when the populations sampled from are leptokurtic, the actual α is less than the nominal α. The opposite is true if platykurtic distributions are sampled from. However, they indicate that these effects are slight. The actual power is less than the nominal power when the populations sampled from are platykurtic, but is greater when the populations sampled from are leptokurtic. These effects can be substantial for small n's.

With these generally fine properties of ANOVA, why then should the Kruskal-Wallis extension and the Normal Scores extension of Levene's test be considered? There are three reasons why:

1. They make much less restrictive assumptions with respect to the type of distribution sampled from.
2. They have been shown to be competitive with the F test with respect to power.

3. They are generally somewhat conservative tests (recall the liberal nature of Levene's test).

These claims will now be documented.

The Kruskal-Wallis technique tests the null hypothesis that the K samples come from identical populations with respect to averages. Valid use of the technique requires that the variable of interest have an underlying continuous distribution, and at least be measured to an ordinal level (8). The Normal Scores tests consider the same hypothesis and make the same assumptions.

It should be noted that there are three forms of the Normal Scores test. They are the Bell-Doksum test, the Terry-Hoeffding test and the Van der Waerden test (8). The Bell-Doksum test has the disadvantage that the test statistic depends upon the particular random normal deviates selected. Two people may both use this test to analyze the same set of data and get different results. The Terry-Hoeffding test requires the use of special tables of expected normal order statistics which are conveniently tabled only for $N \leq 50$. The Van der Waerden test uses the $r/(n+1)$ quantile of a standard normal random variable as the replacement for the score with rank r. The quantiles are widely tabled and can be approximated by interpolation in standard normal tables if desired. This is the type of test which is used in this study.

The relative power of ANOVA, the Kruskal-Wallis test and the normal scores test has been investigated asymptotically and for small samples. Hodges and Lehmann (27) show that the Pitman asymptotic efficiency of the Kruskal-Wallis test with respect to the ANOVA F test is greater than or equal to .864 irrespective of the distribution sampled from. Likewise, the Pitman efficiency of the Normal Scores test with respect to the ANOVA F test is always greater than or equal to 1. They also show that the efficiency of the Kruskal-Wallis test with respect to the normal scores test is somewhere between 0 and $6/\pi = 1.91$, depending on the distribution sampled from. Their proofs are for the two sample case.

While it is reassuring that the Kruskal-Wallis test and the Normal Scores test are asymptotically competitive with the ANOVA F test, the focus of this study is on situations involving small sample sizes.

Klotz (32, p. 631) studied the small sample power and efficiency of the one sample Wilcoxon and Normal Scores test for normal shift alternatives when $N \leq 10$. He concluded: "Because of the extremely high efficiency of the nonparametric tests relative to the t in the region of interest, it is the author's opinion that the nonparametric tests would be preferred to the t in almost all practical situations."

Vander Laan and Oosterhoff (59) compared the Wilcoxon and the normal scores test for the two sample case of equal samples of size six. They sampled from normal distributions and found for various normal-shift alternatives at $\alpha = .05$, that the normal scores test was almost always slightly more powerful than the Wilcoxon test and was always slightly less powerful than the t test. Vander Laan (58) also compared the exact powers using two sample designs with samples of size six when sampling from exponential distributions and also from uniform distributions. In both cases he got results consistent with those obtained when sampling from normal distributions.

Neave and Granger (46, p. 509) when sampling twenty or forty observations from normal or from bimodal asymmetric distributions, concluded:

> Over the range of situations investigated, the normal scores test gave the most satisfactory results, followed closely by the Wilcoxon rank-sum test. Even when the populations were normally distributed, these tests were only slightly inferior to the t test and naturally were much superior in the cases of non-normal populations.

Leaverton and Busch (35) sampling from normal distributions using equal sample sizes of 4, 7, 9, 11, 13 and 25, found the Wilcoxon test compares favorably to the t even for small samples. Using their power curves as reference, they question the widespread use of the ordinary two-sample t test.

McSweeney and Penfield (39) provided tables comparing the power of the Kruskal-Wallis test and the normal scores test for three sample designs with equal sample sizes of 5, 6, 8, 10 and 12. When sampling from uniform distributions, the normal scores test displayed somewhat greater empirical power than the Kruskal-Wallis test. However, when samples were taken from normal distributions, the normal scores test was perhaps slightly less powerful than the Kruskal-Wallis test.

The fact that the Kruskal-Wallis test and the normal scores test are somewhat conservative for small sample situations has been shown in a series of Monte Carlo studies.

Kruskal and Wallis (33) first found their statistic to be slightly conservative for small n. Gabriel and Lachenbruch (13) confirmed this when sampling from three sample designs using various small equal sample sizes. They found the test to be somewhat conservative for almost all of the cases they considered for $\alpha = .10$, .05 and .01.

McSweeney (38) sampled from three sample designs with equal sample sizes of 5, 6, 8, 10 and 12. She indicated that the chi-square approximation was good although slightly conservative for both the

Kruskal-Wallis test and the normal scores test. These results (39, p. 187) seemed to hold when sampling from either normal or uniform distributions. The normal scores test usually was more conservative than was the Kruskal-Wallis test.

Neave and Granger (46, p. 513) observed the same type of results when using normal or bimodal distributions. They found in addition that the normal scores test using inverse scores rounded to one decimal place of accuracy gave results as good as those having four places of accuracy.

Summary

Twelve K-sample tests of scale which have been suggested since 1937 have been examined. These tests have been compared, a few at a time, in recent Monte Carlo studies. With the exception of Box's test and Moses' test, all have been shown to be markedly liberal when sampling from certain types of distributions. For all cases considered Box's test has been shown to be relatively weak when compared to many of the other tests in situations for which the other tests were designed.

Levene's z test was found to be slightly liberal for most distributions considered but was extremely liberal when sampling from a leptokurtic distribution with extreme skew. The test appeared to have satisfactory power with empirical efficiencies relative to Bartlett's test ranging from .83 to .90.

Levene's test involves an ANOVA on the absolute deviations from the sample means. Two other nonparametric tests which are competitors to ANOVA were examined. Both the Kruskal-Wallis test and the Normal Scores test were shown to be competitive with the ANOVA F test with respect to power against shift alternatives while being somewhat more conservative than ANOVA.

When the Kruskal-Wallis test or the Normal Scores test is used in place of Levene's ANOVA technique, perhaps the conservative nature of these nonparametric techniques might counter the liberal nature of Levene's ANOVA to produce a test statistic which has reasonable type I error rates and relatively good power.

CHAPTER III

THE DESIGN

The questions answered by this study were: For the Kruskal-Wallis extension of Levene's test and the Normal Scores extension of Levene's test,

1. How do the nominal type I error rate and the empirical type I error rate correspond?

2. How powerful are these tests with respect to Levene's test?

3. For the two-sample case, how powerful are these tests with respect to the traditional parametric test $F = s_1^2/s_2^2$?

There are a series of concerns which must be addressed when speaking to these questions. Among them are:

1. How many samples should be chosen for each simulation?

2. Which alpha levels will be considered?

3. From which distributions should the X_{ij} come?

4. How many levels of the independent variable will be considered?

5. What should the cell sizes be?

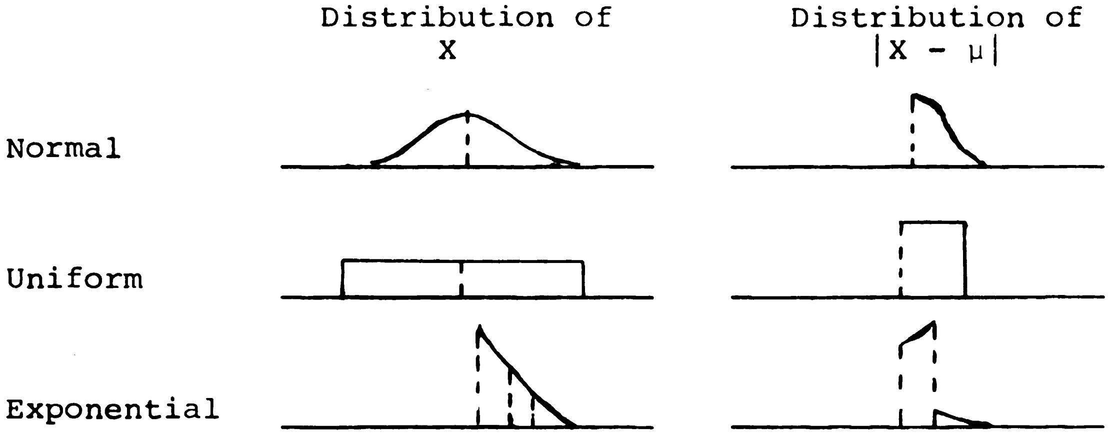
6. Which alternative hypotheses should be selected when looking at relative power?

For each of the cases to be mentioned a simulated analysis was repeated one thousand times and the number of rejections was counted using a series of commonly employed alpha-levels. The alpha levels considered were .10, .05, and .01 since these are the most common levels selected by researchers.

The use of one thousand repetitions somewhat compensates for the disturbing effects of random sampling. The standard errors associated with the empirical estimates of type I error rates are approximately .00949 for $\alpha = .10$, .00689 for $\alpha = .05$ and .00315 for $\alpha = .01$. The standard error associated with a given power estimate is always less than .0158. The approximate value of the error rate estimates and power estimates obtained are therefore reasonably close to the true values.

The distributions considered to test error rates and power were the normal, the uniform and the exponential. It is known that tests for variance are sensitive to nonzero kurtosis. The normal distribution, with zero kurtosis, was selected in order that direct comparisons might be made with established parametric tests. The uniform distribution was selected to represent extreme flatness (platykurtosis) and the exponential distribution was selected to represent extreme peakedness (leptokurtosis) and extreme skew. By the selection of distributions from both ends and the middle of the kurtosis spectrum, the results should apply to most distributions of practical utility in research.

Following are diagrams of the distributions of X and of |X - μ| for the above selections.



Reasonably good fit of nominal type I error rates and empirical type I error rates would be expected if indeed the distribution sampled from was the distribution of absolute deviations from the population mean. In fact the uniform situation has been considered in a somewhat different application in the McSweeney-Penfield (39) study. The empirical estimates were found to be slightly conservative for both the Kruskal-Wallis and the Normal Scores test. Unfortunately, their findings are not applicable in this study since it is concerned with absolute deviations from the sample means, not the population means. The diagrams associated with these $|X - \overline{X}|$ distributions are unknown. Deviating from the sample means rather than the population means, especially when sampling from distributions with a heavy skew such as the exponential, might result in increased variability in the sampling distribution of the statistic. This seems especially likely in situations involving small cell sizes.

Because of the large amount of time used by computers in ranking procedures, the scope of the questions asked was limited to situations where the total N was relatively small. The following situations were considered.

Case No.	1	2	3	4	5	6	7	8	9
No. of Levels of the Independent Variable	2	2	2	3	3	3	4	4	4
Cell size	9	12	18	6	8	12	4	6	9
Total N	18	24	36	18	24	36	16	24	36

The nine cases were selected for the following reasons:

1. Since for each set of levels of the independent variable, the cell sizes increase, the relationships between cell size and power could be examined.

2. Since in cases 1 and 9, cases 2 and 6, and cases 4 and 8, the cell size remains constant as the number of cells increases, some evidence was available to indicate the relationship between the number of cells and the power for a given cell size.

For all nine cases, the normal, uniform and exponential distributions were observed when comparing Levene's test, the Kruskal-Wallis extension of Levene's test and the Normal Scores extension of Levene's test for nominal-empirical error rate agreement and for relative power. In addition, for cases 1, 2 and 3, these three tests were compared with the standard parametric F test where $F = s_1^2/s_2^2$.

Two situations were considered: (a) $\sigma_1 = \sigma_2 = \ldots \sigma_K = 1$ and (b) $\sigma_1 = \sigma_2 = \ldots \sigma_{K-1} = 1$, $\sigma_K = 2$. The first situation permitted study of the nominal-empirical type I error rate question. The second situation gave some evidence as to the relative power of the different tests when the slippage alternative is used. This is an alternative frequently used in Monte Carlo studies and was chosen for convenience. A total of 9 x 3 x 2 = 54 computer simulations of one thousand samples each were made. In order to use computer time efficiently, both situations were observed using the same set of one thousand observations.

The results of the cases involving three cells, in which the uniform distribution was considered, allowed a comparison with the McSweeney-Penfield study. They considered the relative quality of the Kruskal-Wallis test and the Normal Scores test when sampling from normal and uniform distributions.

When computing Levene's statistic and the statistics associated with the two extensions, absolute deviations of observations from their cell means are calculated. If, when sampling from the uniform distribution these deviations were taken from population means, the distribution of deviations would also be uniform. The study of the two extensions would be identical with that of McSweeney and Penfield. However in this study, deviations were taken from sample means and these deviations are not uniformly distributed. By comparison of the two studies the effect of deviating from sample means rather than population means could be observed.

An additional (case 4) exponential situation was run in which the absolute deviations were made from the population means to explore further the effect of high skew and kurtosis on absolute deviations from sample means. One other (case 4) exponential situation was run in which the absolute deviations were made from the sample medians. It was felt that the median would be affected less than the mean by an extreme observation, and thus have less effect on the resulting ranks. The empirical alpha level might coincide more closely with the nominal alpha level. This last experiment was repeated using the normal distribution. Miller's (42) finding that a median deviated Levene-type statistic was asymptotically distribution free provided added incentive to examine the small sample properties of this statistic.

CHAPTER IV

THE GENERATORS USED

In this chapter the different generating procedures used in the study and the tests to which they were subjected are described. The following needed to be generated: (1) pseudorandom numbers between zero and one, (2) uniform random variates, (3) exponential random variates, (4) normal random variates, (5) inverse normal scores and (6) some F values.

Generation of Pseudorandom Numbers

In this study many thousand uniform, exponential and normal random variates needed to be generated. The quality of the procedures in all three of these situations depended upon a program which generated highly dependable pseudorandom numbers.

It has been stated that:

> . . . an acceptable method for generating random numbers must yield sequences of numbers which are (1) uniformly distributed, (2) statistically independent, (3) reproducible and (4) nonrepeating for any desired length. Furthermore, such a method must also be capable of (5) generating random numbers at high rates of speed, yet (6) requiring a minimum amount of computer memory capacity. (45, p. 46)

There were three possible routes that could have been taken to generate pseudorandom numbers: (1) a manual method, (2) the use of a library table or (3) a computer method. Of these three alternatives the only one which was feasible for a study of this magnitude was the use of a computer. The computer used was the IBM 360-30 machine at Ferris State College.

The most common methods used to generate random numbers on a digital computer are called congruential methods, of which three types could have been used: (1) the multiplicative method, (2) the additive method or (3) the mixed method. Since the multiplicative congruential method has been found to behave well statistically (29, p. 240), it was selected for use in this study.

The procedure used in generating random numbers using the multiplicative congruential method involves five steps (45, p. 52).

1. Choose any odd number as a starting value n_o.
2. Choose a value for a. This value should be close to $(2^{b/2} \pm 3)$ where b is the number of bits in the largest possible integer using FORTRAN. With the compiler used, b = 32 so that $(2^{16} + 3)$ was the choice made for a. This choice minimizes the first-order serial correlation between the pseudorandom numbers.

3. Compute $(a \cdot n_i)$ using fixed point integer arithmetic. For the first number generated $i = 0$. The product will consist of $2b = 64$ bits. The 32 low-order bits represent n_{i+1} as the integer multiplication instruction in FORTRAN automatically discards the high order 32 bits.

4. Calculate $r_{i+1} = (n_{i+1})/(2^{32})$ to obtain a uniformly distributed variate on the unit interval.

5. Increment i by 1 and repeat (3) and (4). This cycle is continued until i reaches N, the number of random numbers desired.

Some preliminary testing of the multiplicative congruential generator was performed. To test for uniformity ten thousand numbers were generated by this technique. The numbers generated were sorted into twenty categories of size .05 each. A chi-square test of goodness of fit to the uniform distribution yielded a $\chi^2 = 18.184$ which, when referred to the chi-square distribution with 19 degrees of freedom, indicated no reason to disbelieve that uniformly distributed numbers were being generated $(.50 < p < .75)$.

A serial test to check the degree of randomness between successive numbers in a sequence was run by Sidney Sytsma (56) using one thousand generated numbers. No indication of correlation was found. Work done by

M. D. MacLaren and G. Marsaglia (37) in 1965, and T. E. Hull and A. R. Dobell (30) in 1964 also indicated that the random numbers generated by the multiplicative congruential method are uncorrelated and uniformly distributed.

Close to five million numbers will be generated before a period ends and repetition occurs. This nicely satisfies the fourth criterion. The third and sixth criteria are automatically satisfied by the use of congruential methods since the sequences generated are completely reproducible and require only a minimum amount of computer memory capacity.

Generation of Uniform Random Variates

To generate uniform random variates over the interval (a, b), one simply generates a sequence of random numbers over the interval (0, 1). Each random number (r_i) is then converted to the appropriate number (x_i) on the (a, b) interval by the well-known conversion formula.

$$x_i = a + (b - a) \cdot r_i$$

In this study the intervals of interest were (0, 1) and (0, 2).

Generation of Exponential Random Variates

To generate exponential random variates with mean and standard deviation λ is once again very straight forward. First a sequence of random numbers over the interval (0, 1) is generated. Each random variate (r_i) in the sequence is converted to the appropriate exponential variate by the formula

$$x_i = - \lambda \cdot \log r_i$$

This technique was tested by Sidney Sytsma (56). One thousand exponential variates were generated and shown to have a good fit when compared with the theoretical exponential distribution by means of a chi-square goodness of fit.

Generation of Normal Random Variates

The procedure for simulating normal variates takes the sum of twelve uniformly distributed random variates (r_j, j = 1, 12) where r_j is defined over the interval (0, 1). The Central Limit Theorem guarantees asymptotic convergence to the normal distribution as j becomes large but j = 12 seems to be a good balance between computational efficiency and accuracy. The formula used to obtain a given random variate x_i from a distribution with mean μ and standard deviation σ is

$$x_i = \sigma \cdot [\sum_{j=1}^{12} r_j - 6] + \mu$$

However, this procedure has been found to be unreliable for values of x larger than three standard deviations from the mean. In order to obtain higher accuracy Teichroew devised an approximation technique which improves the accuracy of the tail probabilities (45, p. 93). His technique, which was used in this study, involves as before the generation of twelve uniform random variates to obtain one normal variate. One first computes

$$y_i = \frac{\sum_{j=1}^{12} r_j - 6}{4}$$

and then the variable y is transformed into a standard normal variate z by the formula

$$z_i = a_1 y_i + a_3 y_i^3 + a_5 y_i^5 + a_7 y_i^7 + a_9 y_i^9$$

where:

$a_1 = 3.949846138$

$a_3 = 0.252408784$

$a_5 = 0.076542912$

$a_7 = 0.008355968$

$a_9 = 0.029899776$

To convert the standard normal variate z_i to a normal variate x_i with mean μ and standard deviation σ, simply use the common conversion formula

$$x_i = \sigma \cdot z_i + \mu$$

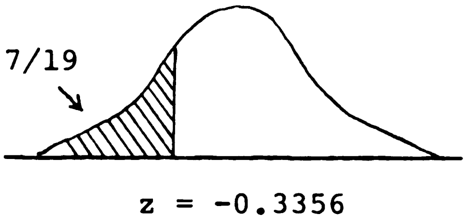
A chi-square goodness of fit test was performed on Teichroew's method of generating random normal deviates by Sidney Sytsma. He generated one thousand normal variates and found the fit to be good (56).

Generation of Inverse Normal Scores

When computing the Normal Scores test, one ranks all observations from 1 to N. These ranks are then treated as percentile points under a normal distribution. Working backwards the z scores can be determined corresponding to the amount of area under the normal curve below these percentile points. Each rank is then replaced by its corresponding z score.

For example, suppose that the total N were 18. The observation with rank 1 would correspond to percentile point 1/19. The observation with rank 2 would correspond to percentile point 2/19. Likewise the observation with rank 18 would correspond to percentile point 18/19.

Now suppose the z score corresponding to percentile point 7/19 is desired. This means that the z value that cuts off the lowermost 7/19 of the normal curve is desired. This z value is -0.3356. So the observation



which is ranked 7 would employ the value -0.3356 when the Normal Scores test statistic is computed. All other z values are obtained in this manner.

Many of the z values necessary for use in this study were not previously tabled as accurately as would be desired. Consequently, a procedure described by C. Hastings, Jr. (25) was used to derive these values from the ranks. This approximation is correct to within 0.0004 units of the true inverse normal value.

Let u be the percentile point that is to be replaced by an inverse normal score where $u = r/(n+1)$. When $0 < u \leq .5$, let $z = \sqrt{-2 \ln u}$. Then

$$y = -\left(z - \frac{a_o + a_1 z + a_2 z^2}{1 + b_1 z + b_2 z^2 + b_3 z^3} \right)$$

is the desired approximation for the inverse of the standard normal distribution where

$a_o = 2.515517$ $b_1 = 1.432788$

$a_1 = 0.802853$ $b_2 = 0.189269$

$a_2 = 0.010328$ $b_3 = 0.001308$

When $u > .5$, let $z = \sqrt{-2 \ln (1 - u)}$. Use

$$y = z - \frac{a_o + a_1 z + a_2 z^2}{1 + b_1 z + b_2 z^2 + b_3 z^3}$$

with the constants a_j ($j = 0, 1, 2$) and b_j ($j = 1, 2, 3$) defined as before to obtain the desired approximation.

Generation of F Values

When the 2-sample variance ratio test or Levene's Test is used, the sample test statistic is compared with a value of the F distribution with m and n degrees of freedom. For many values of m and n, tabled values of the F distribution are readily available. In this study the available values were used whenever possible; however, many of the m and n pairings necessary were not tabled for the α level desired.

A computer program to obtain either F values or their associated probabilities was written in FORTRAN by Clark Holloway and W. B. Capp in 1959 and revised by R. J. McKelvey in 1961 (28). Its use was suggested by Linda Glendening (20). This program works in either of two directions: if m, n and F are entered, the corresponding probability value p is calculated; if m, n and p are entered, the corresponding F value is calculated.

The second option was desirable in this study, however, the time taken to calculate a given F when running in background on the IBM 360-30 computer was close to 30 minutes in the three runs attempted. The compute time to run in the first direction was found to be many times faster than that of the second.

Consequently, a main program was written in which values of m, n and an F which was known to be somewhat smaller than the desired F value were entered. The Holloway-Capp-McKelvey program was used as a subroutine. The entered F value was repeatedly incremented in steps of .10 until a p higher than desired was obtained. At this point, the last low estimate was incremented in steps of .01. The procedure was repeated with the final increment of .001. This gave F value estimates to three decimal places. To test the quality of this procedure, four known tabled values of F were subjected to it. In all cases the calculated F value was identical to the value found in the tables.

CHAPTER V

THE RESULTS

The basic questions of this study are concerned with the relative quality of four different tests of homogeneity of variance: (1) the standard F ratio, (2) Levene's test, (3) a Kruskal-Wallis extension of Levene's test and (4) the Normal Scores extension of Levene's test. Which test has the best correspondence between nominal type I error rate and empirical type I error rate? Which test has the best ability to correctly reject the null hypothesis when it is appropriate to do so?

Three different distributions were considered and for each distribution comparisons were made between nominal and empirical type I error rates. The nominal error rates used were $\alpha = .10$, $\alpha = .05$ and $\alpha = .01$. The distributions considered were the normal, the uniform and the exponential distribution.

In this chapter, the results of the study are reported first for the normal distribution, then for the uniform distribution and finally for the exponential distribution. These results suggested a modification using absolute deviations from the median rather than the mean. Two simulations were performed using this modification; the results of these simulations are reported. The chapter ends with a short summary of the results.

Sampling from Normal Distributions

The results of the simulation using random samples from the normal distribution are presented in two tables. Displayed in Table 5-1 is the empirical type I error rate for the four tests using one thousand random samples from normal distributions with $\sigma_1 = \sigma_2 = . . . \sigma_K = 1$. The standard errors associated with the estimates for the null case are approximately .00949 for $\alpha = .10$, .00689 for $\alpha = .05$ and .00315 for $\alpha = .01$. The power of the four tests using random samples from normal distributions with $\sigma_1 = \sigma_2 = . . . \sigma_{K-1} = 1$, $\sigma_K = 2$, is displayed in Table 5-2. The size of the standard error associated with a given power estimate is always less than .0158.

It can be observed (Table 5-2) that, for a fixed number of cells, as the cell size increased the power significantly increased.

For a fixed cell size as the number of cells increased, Levene's test always became significantly more powerful. This relationship did not seem to hold

TABLE 5-1.--Empirical estimates of type I error rates using the F test, Levene's test, the Kruskal-Wallis extension and the Normal Scores extension when sampling from normal distributions with $\sigma_1 = \sigma_2 = . . . \sigma_K = 1$.

Cell Size	α = .100				α = .050				α = .010			
	F	L	K	N	F	L	K	N	F	L	K	N
9, 9	.101	.100	.091	.098	.050	.056	.046	.043	.014	.013	.011	.010
12, 12	.102	.110	.113	.104	.049	.059	.056	.053	.010	.013	.008	.008
18, 18	.104	.116	.118	.115	.052	.063	.062	.059	.012	.010	.013	.012
6, 6, 6	-	.134	.143	.137	-	.080	.069	.064	-	.017	.007	.005
8, 8, 8	-	.133	.118	.116	-	.066	.059	.061	-	.018	.015	.011
12, 12, 12	-	.093	.096	.094	-	.050	.053	.046	-	.009	.007	.006
4, 4, 4, 4,	-	.160	.151	.140	-	.101	.082	.074	-	.041	.007	.003
6, 6, 6, 6	-	.148	.135	.136	-	.080	.064	.060	-	.020	.005	.004
9, 9, 9, 9	-	.107	.092	.090	-	.053	.055	.054	-	.014	.008	.006

TABLE 5-2.--Empirical estimates of power using the F test, Levene's test, the Kruskal-Wallis extension and the Normal Scores extension when sampling from normal distributions with $\sigma_1 = \sigma_2 = . . . \sigma_{K-1} = 1, \sigma_K = 2$

Cell Size	$\alpha = .100$				$\alpha = .050$				$\alpha = .001$			
	F	L	K	N	F	L	K	N	F	L	K	N
9, 9	.650	.561	.473	.493	.515	.384	.344	.333	.254	.119	.109	.102
12, 12	.751	.674	.561	.578	.660	.494	.437	.418	.411	.201	.183	.169
18, 18	.917	.853	.734	.739	.849	.738	.609	.611	.658	.427	.351	.342
6, 6, 6	-	.450	.357	.353	-	.296	.215	.224	-	.119	.070	.066
8, 8, 8	-	.523	.389	.405	-	.366	.273	.279	-	.160	.095	.096
12, 12, 12	-	.719	.581	.596	-	.601	.428	.465	-	.331	.204	.221
4, 4, 4, 4,	-	.378	.273	.275	-	.233	.167	.158	-	.091	.009	.006
6, 6, 6, 6	-	.457	.311	.331	-	.321	.199	.206	-	.137	.054	.055
9, 9, 9, 9	-	.625	.456	.482	-	.516	.323	.355	-	.282	.121	.133

for either extension of Levene's test as sometimes the power increased and sometimes it decreased with an increasing number of cells. When considering the extensions, 83 percent of these increases or decreases were within the .95 confidence interval of ordinary sampling variability.

The average empirical estimates of α are presented in Table 5-3. Averages over the various cell sizes were used to condense the larger tables to a manageable size. By averaging, the extremes become somewhat obscured. First consider the two cell designs. The three alternatives to the standard F test appear to be somewhat liberal in their estimates of nominal α. For all three nominal α-levels considered, Levene's test was the most liberal with the Normal Scores extension the least liberal.

The standard F test not only gives the best empirical estimate of type I error rates, it also has the highest power (Table 5-4). Levene's test is more powerful than either of the extensions, which are approximately equally powerful.

To compare Levene's test and its two extensions further, consider the average empirical estimates of type I error rates for the three- and four-cell designs (Table 5-3). As in the two cell designs, Levene's test is the most liberal and the Normal Scores extension the

TABLE 5-3.--Average empirical estimates of type I error rates for the four tests considered when sampling from normal distributions

	Nominal α	F Test	Levene's Test	Kruskal-Wallis Extension	Normal Scores Extension
	.100	.102	.109	.107	.106
2 Cells	.050	.050	.059	.055	.052
	.010	.012	.012	.011	.010
	.100	-	.120	.119	.116
3 Cells	.050	-	.065	.060	.057
	.010	-	.015	.010	.007
	.100	-	.138	.126	.122
4 Cells	.050	-	.078	.067	.063
	.010	-	.025	.007	.004

TABLE 5-4.--Average empirical estimates of power for the four tests considered when sampling from normal distributions

	Nominal α	F Test	Levene's Test	Kruskal-Wallis Extension	Normal Scores Extension
	.100	.773	.696	.589	.603
2 Cells	.050	.675	.539	.463	.454
	.010	.441	.249	.214	.204
	.100	-	.564	.442	.451
3 Cells	.050	-	.421	.305	.323
	.010	-	.203	.123	.128
	.100	-	.487	.347	.363
4 Cells	.050	-	.357	.230	.240
	.010	-	.170	.061	.065

least liberal. Poorer empirical estimates of type I error rate occur with increasing cell size.

In all, twenty-seven situations have been studied (Table 5-1). Fifteen times the Normal Scores extension is closer to nominal α than is the Kruskal-Wallis extension, and eighteen times it is closer than Levene's test. Twenty-three times the Normal Scores extension yielded a lower estimate of α than did the Kruskal-Wallis extension and twenty-three times it was lower than Levene's test. The average distance the empirical estimate of α was from its expected value can be observed in Table 5-5. The empirical estimate of α provided by the Normal Scores Extension is, on the average, closer to the nominal α than are those of the other tests. Levene's test comes in a poor third.

TABLE 5-5.--Average distance of the empirical estimate of α from the expected value of α for all designs considered when sampling from normal distributions

Nominal α	Levene's Test	Kruskal-Wallis Extension	Normal Scores Extension
.100	.024	.022	.018
.050	.018	.012	.010
.010	.007	.003	.003

As in the two-cell situation, Levene's test is found to be the most powerful of the three tests considered (Table 5-4) for three- and four-cell designs.

However, some of the advantage with respect to power may be attributable to the slightly greater liberality of Levene's test. The Normal Scores extension was more powerful than the Kruskal-Wallis extension for all α levels considered.

Sampling from Uniform Distributions

The results of the simulation using random samples from the uniform distribution are presented in two tables. In Table 5-6 the empirical type I error rate for the four tests, using random samples from uniform distributions with $\sigma_1 = \sigma_2 = . . . = \sigma_K = 1$, is displayed. The power of the four tests, using random samples from uniform distributions with $\sigma_1 = \sigma_2 = . . . = \sigma_{K-1} = 1$, $\sigma_K = 2$, is displayed in Table 5-7.

It can be observed that, for a fixed number of cells, as the cell size increased the power significantly increased. For a fixed cell size as the number of cells increased, Levene's test usually became significantly more powerful. The direction of this relationship usually held for both extensions of Levene's test although the size of the increase was usually insignificant.

Once again, consider the two-cell designs. The average empirical estimates of α are presented in Table 5-8. The values of the empirical estimates of

TABLE 5-6.--Empirical estimates of type I error rates using the F test, Levene's test, the Kruskal-Wallis extension and the Normal Scores extension when sampling from uniform distributions with $\sigma_1 = \sigma_2 = . . . \sigma_K = 1$

Cell Size	α = .100				α = .050				α = .010			
	F	L	K	N	F	L	K	N	F	L	K	N
9, 9	.038	.143	.137	.138	.017	.081	.079	.077	.003	.019	.016	.012
12, 12	.024	.111	.121	.118	.008	.063	.068	.056	.000	.016	.011	.009
18, 18	.021	.121	.127	.117	.003	.065	.064	.058	.000	.010	.008	.006
6, 6, 6	-	.143	.133	.131	-	.077	.064	.059	-	.021	.009	.010
8, 8, 8	-	.123	.137	.128	-	.077	.069	.067	-	.014	.006	.005
12, 12, 12	-	.123	.114	.106	-	.053	.058	.050	-	.007	.005	.004
4, 4, 4, 4	-	.169	.178	.161	-	.112	.090	.079	-	.041	.005	.003
6, 6, 6, 6	-	.149	.142	.134	-	.089	.071	.063	-	.021	.010	.010
9, 9, 9, 9	-	.141	.139	.123	-	.072	.068	.064	-	.016	.010	.011

TABLE 5-7.--Empirical estimates of power using the F test, Levene's test, the Kruskal-Wallis extension and the Normal Scores extension when sampling from uniform distributions with $\sigma_1 = \sigma_2 = . . . \sigma_{K-1} = 1, \sigma_K = 2$

Cell Size	α = .100				α = .050				α = .010			
	F	L	K	N	F	L	K	N	F	L	K	N
9, 9	.625	.640	.496	.517	.423	.479	.388	.387	.117	.237	.179	.160
12, 12	.780	.796	.669	.680	.614	.661	.568	.550	.260	.379	.264	.251
18, 18	.946	.923	.833	.827	.884	.863	.733	.736	.595	.650	.502	.496
6, 6, 6	-	.513	.422	.427	-	.386	.262	.281	-	.173	.085	.074
8, 8, 8	-	.653	.519	.520	-	.535	.369	.390	-	.277	.162	.167
12, 12, 12	-	.816	.656	.686	-	.737	.540	.566	-	.516	.300	.310
4, 4, 4, 4	-	.365	.329	.322	-	.260	.211	.200	-	.134	.017	.011
6, 6, 6, 6	-	.531	.406	.420	-	.429	.270	.291	-	.231	.092	.089
9, 9, 9, 9	-	.721	.547	.588	-	.621	.410	.458	-	.398	.191	.211

TABLE 5-8.--Average empirical estimates of type I error rates for the four tests considered when sampling from uniform distributions

	Nominal α	F Test	Levene's Test	Kruskal-Wallis Extension	Normal Scores Extension
2 Cells	.100	.028	.125	.128	.124
	.050	.009	.070	.070	.064
	.010	.001	.015	.012	.009
3 Cells	.100	-	.130	.128	.122
	.050	-	.069	.064	.059
	.010	-	.014	.007	.006
4 Cells	.100	-	.153	.153	.139
	.050	-	.091	.076	.069
	.010	-	.026	.008	.008

type I error rates tend to be extremely conservative for the F test. The other three tests yield somewhat liberal estimates of the type I error rates. When samples are drawn from uniform distributions, the Normal Scores extension gave better estimates of nominal α than either Levene's test or the Kruskal-Wallis extension. This is consistent with the results obtained when sampling from the normal distribution.

When samples are drawn from uniform distributions using two-cell designs, Levene's test appears to be somewhat more powerful than the F test (Table 5-9). The difference is especially noticeable at $\alpha = .010$, where the F test is extremely conservative and Levene's test is somewhat liberal. The Kruskal-Wallis extension and the Normal Scores extension both have approximately the same power, but both extensions are less powerful than Levene's test.

For the three- and four-cell designs, the average empirical estimates of type I error rates for Levene's test and the two extensions are also presented in Table 5-8. Twenty-three times out of twenty-seven the Normal Scores extension gave a lower estimate of nominal α than did the Kruskal-Wallis extension. Twenty-five times the Normal Scores estimate was lower than that of Levene's test. As in all previous situations examined in this study, Levene's test appears to be the most liberal with

TABLE 5-9.--Average empirical estimates of power for the four tests considered when sampling from uniform distributions

	Nominal α	F Test	Levene's Test	Kruskal-Wallis Extension	Normal Scores Extension
2 Cells	.100	.784	.786	.666	.675
	.050	.640	.668	.563	.558
	.010	.324	.422	.315	.302
3 Cells	.100	-	.661	.532	.544
	.050	-	.553	.390	.412
	.010	-	.322	.182	.184
4 Cells	.100	-	.539	.427	.443
	.050	-	.437	.297	.316
	.010	-	.254	.100	.104

the Normal Scores extension the least. Nineteen times out of twenty-seven the Normal Scores extension estimate was closer to nominal α than the Kruskal-Wallis extension estimate. Twenty-two times it was closer than the Levene estimate.

It is clear that for the situations involving sampling from the uniform distribution, the Normal Scores extension provides better estimates of nominal α than does either Levene's test or the Kruskal-Wallis extension. This is also borne out when considering the average distance that the empirical estimates are from the nominal α (Table 5-10).

TABLE 5-10.--Average distance of the empirical estimates of α from the expected value of α for all designs considered when sampling from uniform distributions

Nominal α	Levene's Test	Kruskal-Wallis Extension	Normal Scores Extension
.100	.036	.036	.028
.050	.027	.020	.014
.010	.009	.003	.003

As happened when samples were chosen from normal distributions using three- and four-cell designs, Levene's test was the most powerful of the three tests considered (Table 5-9), and the Normal Scores extension was slightly more powerful than the Kruskal-Wallis extension.

The empirical type I error rates obtained when sampling from the uniform distribution for the Kruskal-Wallis extension and the Normal Scores extension can be observed in Table 5-11. Both statistics were calculated on the basis of observations deviated from their respective populations means and from sample means. When deviations were taken from the sample means, the estimates of type I error rates for α = .10 and α = .05 were quite high. This indicated that the sample mean is not a robust estimator of the population mean, especially for the smaller sample sizes.

TABLE 5-11.--Empirical estimates of type I error rates obtained when sampling from uniform distributions for the Kruskal-Wallis extension and the Normal Scores extension. Deviations are taken from (A) population means* and (B) sample means

Cell Size		α = .10		α = .05		α = .01	
		K	N	K	N	K	N
6, 6, 6	A	.084	.084	.037	.036	.004	.004
	B	.133	.131	.064	.059	.009	.010
8, 8, 8	A	.095	.094	.045	.041	.006	.005
	B	.137	.128	.069	.067	.006	.005
12, 12, 12	A	.107	.102	.047	.045	.009	.008
	B	.114	.106	.058	.050	.005	.004

*Obtained from McSweeney-Penfield (38, p. 187) study.

Sampling from Exponential Distributions

The results of the simulation using random samples from the exponential distribution are presented in two tables. Displayed in Table 5-12 is the empirical type I error rate for the four tests using one thousand random samples from exponential distributions with $\sigma_1 = \sigma_2 = \ldots = \sigma_K = 1$. The power of the four tests using random samples from exponential distributions with $\sigma_1 = \sigma_2 = \ldots = \sigma_{K-1} = 1$, $\sigma_K = 2$, is displayed in Table 5-13.

For all four tests used, the empirical estimates of α are very liberal. The poor nature of the fit is summarized in Table 5-14. When 100 rejections are expected, the closest any test comes is 243 rejections. With 50 expected, the closest is 151 and when 10 rejections are expected, the closest is 51. With a fit this poor, it makes little sense to consider the question of relative power.

Agreement Rates of Kruskal-Wallis and Normal Scores Extensions

So far gross comparisons of the empirical type I error rates have been made across tests and for each of three distributions. The availability of one extensive set of tests results permitted consideration of the issue how the Kruskal-Wallis and Normal Scores extensions dealt with the same data set.

TABLE 5-12.--Empirical estimates of type I error rates using the F test, Levene's test, the Kruskal-Wallis extension and the Normal Scores extension when sampling from exponential distributions with $\sigma_1 = \sigma_2 = . . . \sigma_K = 1$

Cell Size	α = .100				α = .050				α = .010			
	F	L	K	N	F	L	K	N	F	L	K	N
9, 9	.308	.243	.286	.287	.211	.151	.195	.178	.093	.055	.065	.051
12, 12	.347	.262	.311	.292	.254	.172	.219	.198	.126	.053	.078	.067
18, 18	.358	.230	.316	.293	.265	.148	.233	.214	.143	.053	.098	.086
6, 6, 6	-	.337	.357	.339	-	.210	.238	.227	-	.093	.083	.071
8, 8, 8	-	.285	.341	.324	-	.168	.234	.223	-	.066	.081	.074
12, 12, 12	-	.305	.390	.386	-	.210	.296	.273	-	.076	.120	.099
4, 4, 4, 4	-	.449	.448	.425	-	.270	.279	.264	-	.077	.033	.018
6, 6, 6, 6	-	.386	.428	.411	-	.258	.315	.286	-	.107	.102	.094
9, 9, 9, 9	-	.342	.449	.428	-	.243	.318	.288	-	.087	.128	.109

TABLE 5-13.--Empirical estimates of power using the F test, Levene's test, the Kruskal-Wallis extension and the Normal Scores extension when sampling from exponential distributions with $\sigma_1 = \sigma_2 = . . . \sigma_{K-1} = 1, \sigma_K = 2$

Cell Size	α = .100				α = .050				α = .010			
	F	L	K	N	F	L	K	N	F	L	K	N
9, 9	.555	.464	.486	.477	.466	.337	.392	.361	.302	.155	.202	.166
12, 12	.630	.564	.590	.570	.558	.433	.496	.476	.405	.216	.285	.247
18, 18	.749	.668	.714	.679	.671	.557	.614	.586	.511	.332	.424	.395
6, 6, 6	-	.495	.519	.496	-	.342	.378	.359	-	.171	.147	.134
8, 8, 8	-	.528	.578	.541	-	.403	.433	.413	-	.204	.223	.211
12, 12, 12	-	.617	.655	.642	-	.498	.550	.527	-	.300	.346	.322
4, 4, 4, 4	-	.508	.501	.484	-	.353	.359	.341	-	.144	.054	.035
6, 6, 6, 6	-	.546	.558	.543	-	.419	.424	.417	-	.211	.204	.179
9, 9, 9, 9	-	.602	.651	.620	-	.476	.537	.520	-	.271	.293	.266

TABLE 5-14.--Average distance of the empirical estimates of α from the expected value of α for all designs considered when sampling from exponential distributions

Nominal α	Levene's Test	Kruskal-Wallis Extension	Normal Scores Extension	F Test*
.100	.315	.369	.354	.336
.050	.203	.259	.239	.243
.010	.074	.088	.074	.171

*Only two-cell designs considered.

The number of agreements of the Kruskal-Wallis and Normal Scores extensions can be observed in Table 5-15 for the situation in which the null hypothesis is true and α = .10. The rate of agreement lies between 95.4 percent and 98.9 percent. The same type of information is displayed in Table 5-16 for the situation in which the alternative hypothesis is true. In this case, the rate of agreement lies between 94.3 percent and 97.8 percent.

A Modification

The extremely liberal situation which existed when sampling from the exponential distribution strongly suggests that deviating from the sample means does not produce the most desirable results. The effect a single large value can have on the sample mean produces an instability which strongly affects the deviations.

TABLE 5-15.--Kruskal-Wallis extension and Normal Scores extension agreement rates for one thousand trials* (α = .10, $\sigma_1 = \sigma_2 = . . . \sigma_K = 1$)

Decisions Cited as	Normal				Uniform				Exponential			
(K-W,N.S.)	(A,A)	(R,R)	(A,R)	(R,A)	(A,A)	(R,R)	(A,R)	(R,A)	(A,A)	(R,R)	(A,R)	(R,A)
Cell Size												
6, 6, 6	849	129	8	14	856	120	11	13	878	76	21	25
8, 8, 8	873	107	9	11	862	127	1	10	651	316	8	25
12, 12, 12	895	85	9	11	876	96	10	18	594	370	16	20
4, 4, 4, 4	846	137	3	14	820	159	2	19	545	418	7	30
6, 6, 6, 6	848	119	17	16	851	127	7	15	565	404	7	24
9, 9, 9, 9	901	83	7	9	856	118	5	21	541	418	10	31

*A = accept H_o; R = reject H_o

TABLE 5-16.--Kruskal-Wallis extension and Normal Scores extension agreement rates for one thousand trials* ($\alpha = .10$, $\sigma_1 = \sigma_2 = . . . \sigma_{K-1} = 1$, $\sigma_K = 2$)

Decisions Cited as	Normal				Uniform				Exponential			
(K-W,N.S.)	(A,A)	(R,R)	(A,R)	(R,A)	(A,A)	(R,R)	(A,R)	(R,A)	(A,A)	(R,R)	(A,R)	(R,A)
Cell Size												
6, 6, 6	629	339	14	18	555	404	23	18	473	488	8	31
8, 8, 8	579	373	32	16	461	500	20	19	420	539	2	39
12, 12, 12	389	566	30	15	305	647	39	9	338	635	7	20
4, 4, 4, 4	715	263	12	10	660	311	11	18	490	475	9	26
6, 6, 6, 6	654	296	35	15	568	394	26	12	434	535	8	23
9, 9, 9, 9	511	449	33	7	404	539	49	8	342	613	7	38

*A = accept H_o; R = reject H_o

This point can be demonstrated through the use of a numerical example. Suppose a sample of size four is taken from a highly skewed distribution. Suppose the first three observations are 0.2, 0.5 and 0.8. If the fourth observation were 1.3, then the mean would be 0.7 and the absolute mean deviations would be 0.5, 0.2, 0.1 and 0.6. If the fourth observation were 3.3, then the mean would be 1.2 and the absolute mean deviations would be 1.0, 0.7, 0.4 and 2.1. It is clear that the first three deviations were strongly affected by the fourth observation; however, if these observations were median deviated, the first three deviations would be unaffected.

To study the statistics in the absence of this extreme instability, a run was made in which deviations were taken from the population means rather than the sample means. This run produced slightly conservative estimates of the nominal α (Table 5-17).

In exploring the asymptotic properties of Levene's test, Miller (42) concluded that if deviations were taken from the sample means, Levene's test statistic would not be asymptotically distribution free. However, he was able to prove that it would be asymptotically distribution free if deviations were taken from sample medians.

TABLE 5-17.--Empirical type I error rate and power obtained by three variations of Levene's test and the Kruskal-Wallis and Normal Scores extensions where a three-cell design (6, 6, 6) was used when sampling from exponential distributions with σ_1, σ_2 and σ_3 as tabled

Deviations From:	σ_1	σ_2	σ_3	α = .100 L	α = .100 K	α = .100 N	α = .050 L	α = .050 K	α = .050 N	α = .010 L	α = .010 K	α = .010 N
Population	1	1	1	.090	.101	.097	.046	.045	.042	.005	.005	.005
Means	1	1	2	.379	.394	.381	.249	.244	.233	.096	.064	.054
Sample	1	1	1	.337	.357	.339	.210	.238	.227	.093	.083	.071
Means	1	1	2	.495	.519	.496	.342	.378	.359	.171	.147	.134
Sample	1	1	1	.107	.189	.183	.054	.107	.103	.027	.019	.017
Medians	1	1	2	.216	.328	.328	.135	.211	.205	.043	.060	.043

For this reason and the fact an outlier has much less effect on the median than on the mean, another run was made to see if absolute deviations from the median might prove promising. The exponential distribution was used since this is the situation in which all previous attempts had failed. The results of this run are also shown in Table 5-17.

From this one run, it appears that deviating from the sample medians rather than the sample means produces much better fit of empirical α to the nominal α. This is especially true for Levene's test at the nominal α levels .10 and .05. However, the fit at $\alpha = .01$ remains poor even though considerably improved from the former case. The power is substantially reduced, even relative to the more conservative population mean deviated scores, when median deviation is used. Hence, the loss of power is not totally a consequence of obtaining a better actual alpha level.

To see if this technique would perform well when sampling from other distributions, a final run was made in which samples were taken from normal distributions. The results are reported in Table 5-18. Once again, the Levene type test performed well. For $\alpha = .05$ and $\alpha = .10$, the empirical estimates of type I error rates are no longer liberal. However, as in the case of

TABLE 5-18.--Empirical type I error rate and power obtained by two variations of Levene's test and the Kruskal-Wallis and Normal Scores extensions where a three-cell design (6, 6, 6) was used when sampling from normal distributions with σ_1, σ_2 and σ_3 as tabled

Deviations From:				α = .100			α = .050			α = .010		
	σ_1	σ_2	σ_3	L	K	N	L	K	N	L	K	N
Sample	1	1	1	.134	.143	.137	.080	.069	.064	.017	.007	.005
Means	1	1	2	.450	.357	.353	.296	.215	.224	.119	.070	.066
Sample	1	1	1	.093	.122	.135	.049	.068	.074	.014	.009	.005
Medians	1	1	2	.296	.298	.311	.191	.175	.180	.064	.055	.047

sampling from exponential distributions, the estimates are not as close if $\alpha = .01$.

Summary of Results

Three questions were of concern in this study. The first question was: For the Kruskal-Wallis extension of Levene's test and the Normal Scores extension of Levene's test, how do the nominal type I error rate and the empirical type I error rate correspond?

When sampling from normal or uniform distributions with nominal α's of .10 and .05, both extensions gave slightly liberal estimates of the nominal α. The Normal Scores extension consistently gave the better estimate. When the nominal α was .01, the extensions gave, on the average, slightly conservative estimates. At this α level, they seem to have nearly the same quality.

The strongest criticism of Levene's test has been its liberal nature. Both extensions are less liberal in their observed type I error rates than Levene's test.

When the exponential distribution was considered, all observed type I error rates were extremely high, no matter which of the four tests was used. This showed the undesirable effect an outlier can have on a sample mean.

Since outliers have little effect on a median, two runs were made in which deviations were made from sample medians rather than sample means. When exponential distributions or normal distributions were used with this technique, a good fit was found for the Levene type test when using the nominal α's of .10 and .05. However, the technique was also too liberal for $\alpha = .01$.

The second question to be addressed in this study was: how powerful are the Kruskal-Wallis extension and the Normal Scores extension with respect to Levene's test? In view of the poor observed type I error rates for the exponential distributions, the power comparison was made only when sampling from normal or uniform distributions. When deviating from sample means, the Normal Scores extension was found in most cases to be slightly more powerful than the Kruskal-Wallis extension. Both extensions were less powerful than Levene's test.

The third question of concern was: for the two sample case, how powerful are these extensions with respect to the traditional F test? Although both extensions were more stable across distributions in the observed type I error rate than the traditional variance ratio F test, they have somewhat less power, at least when sampling from normal or uniform distributions.

CHAPTER VI

SUMMARY, CONCLUSIONS AND DISCUSSION

Summary

This study was motivated by the need for a good K-sample test for equal variability. Well over a dozen tests have been suggested in the literature but all suffer from one or more of the following ailments:

(1) Poor correspondence between nominal type I error rate and empirical type I error rate if the normality assumption is violated;

(2) Low power when this correspondence is good;

(3) Lack of post hoc procedures associated with the method.

Of all the tests which have been considered in the recent literature, only two have shown an acceptable correspondence between the nominal and empirical type I error rates, no matter what distribution is sampled from. These are Box's test and Moses' test. Box's test was found to be more powerful than Moses' test for all cases considered. Unfortunately, neither of these procedures is desirable for sample sizes much less than fifteen.

This is because both tests involve breaking each of the K samples into subsamples and computing ln s^2 for each subsample.

Of the remaining tests, one suggested by Levene showed the most promise. Levene's test consists of an analysis of variance on the absolute deviations of the observations from their sample means. The average amount by which the empirical and nominal type I error rates differed was relatively low for this test; however, Levene's test consistently gave somewhat liberal estimates of the nominal alpha. Both the Kruskal-Wallis test and the Normal Scores test of location are known to be somewhat more conservative than is ANOVA for most distributions sampled. Usually little power is sacrificed when performing the Kruskal-Wallis test or the Normal Scores test. It was thought that the conservative nature of these tests might counter the liberal nature of Levene's test to produce a test statistic which was acceptable in its empirical type I error rate. Both the Kruskal-Wallis extension and the Normal Scores extension of Levene's test would hopefully have power comparable to that of Levene's test. Thus, the thrust of this study was a Monte Carlo investigation of the properties of both the Kruskal-Wallis extension and the Normal Scores extension of Levene's test for equal variability.

Two, three and four levels of the independent variable were considered for various small sample sizes. The properties of the test statistics were observed when sampling from normal, uniform and exponential distributions. For each of the cases a simulated analysis was repeated one thousand times and the number of rejections of the null hypothesis of equal variability was counted using a series of commonly employed nominal alpha levels. Both nominal-empirical alpha level fit and power were of concern in this study.

When samples were taken from either uniform or normal distributions, the results were in the direction predicted. Both extensions of Levene's test produced better estimates of type I error rate than did Levene's test. The Normal Scores extension proved to give generally better estimates of type I error rate and power than did the Kruskal-Wallis extension. For $\alpha = .10$ and $\alpha = .05$, both extensions were still somewhat liberal and for $\alpha = .01$ they were slightly conservative. In these same situations, Levene's test had slightly more power; however, some of the advantage with respect to power may be attributable to the slightly greater liberality of Levene's test.

When the exponential distribution was sampled from, the empirical estimates of α were all very liberal. With this poor quality fit, the power comparisons were meaningless. With the failure of these techniques for the exponential distribution, the good K-sample test for equal variability which is somewhat distribution free had not been found. This failure indirectly suggested the use of a similar but slightly different technique.

The properties of a modified form of Levene's test and of the two extensions were observed in a mini-study. The test statistics were computed as before, with the exception of deviating observations from sample medians rather than sample means. The median deviation technique was employed for the three-sample design with six observations per sample when sampling from normal distributions and when sampling from exponential distributions. The results of this mini-study were quite encouraging.

For $\alpha = .10$ and $\alpha = .05$, the empirical type I error rates of the Levene type test were quite close to the nominal level when sampling from either distribution. However, for $\alpha = .01$ the empirical type I error rates were too liberal. With the exception of Box's test and Moses' test, this is the only time that the type I error rate was well behaved when sampling from either a distribution with high kurtosis or with heavy skew. Although the Levene type test using the median was not without power, it appears to be less powerful than the appropriate conventional tests when sampling from normal distributions. The Kruskal-Wallis median extension and the Normal Scores median extension of the Levene type test did not fare nearly as well in either their empirical estimates of type I error rate or power.

Findings

The first seven findings result from the main study; the last two findings are from a mini-study performed as a follow-up to the main study. The latter must be considered tentative in light of the limited nature of the mini-study.

When sampling from normal distributions or uniform distributions:

1. Both the Kruskal-Wallis extension and the Normal Scores extension of Levene's test gave better estimates of type I error rates than did Levene's test.

2. In general the Normal Scores extension gave slightly better estimates of type I error rates than did the Kruskal-Wallis extension.

3. Both extensions were generally somewhat liberal in their estimates of type I error rates at $\alpha = .10$ and $\alpha = .05$ and somewhat conservative for $\alpha = .01$.

4. Although Levene's test was more powerful than either extension, a part of this extra power came from the liberal nature of the test.

5. The Normal Scores extension was generally slightly more powerful than was the Kruskal-Wallis extension.

6. The two sample F ratio was more powerful than Levene's test and its extensions when sampling from normal distributions. However, when sampling from the uniform distributions, Levene's test was more powerful than the F ratio.

When sampling from exponential distributions:

7. Levene's test and the two extensions all gave considerably liberal estimates of type I error rates.

When deviating from the sample medians rather than the sample means, and when sampling from normal or exponential distributions:

8. The Levene type test gave good estimates of type I error rates for $\alpha = .10$ and $\alpha = .05$ but was somewhat liberal for $\alpha = .01$.

9. The two extensions gave poorer estimates of type I error rates than did the Levene test and they were also less powerful.

Discussion

This study was originally prompted by the practical need to test a hypothesis of equal variability in a design which had several independent variables. No general multi-factor tests that were insensitive to non-normality could be identified, and those single factor tests which were identified proved to be defective in one or more ways. Nevertheless, some guidance can be given to the researcher who needs a pragmatic answer to the question of how to proceed with his data analysis.

If the sample sizes are reasonably large and of equal size, the researcher can treat the study design as if it consisted of a single factor and can use Box's test. It has been shown to give reasonably good estimates of type I error rate for a variety of distributions when the sample sizes are at least fifteen. This test has post hoc procedures available since it is an ANOVA of $\ln s^2$. Its primary drawback appears to be the sacrifice in power necessitated by the grouping of observations to form the $\ln s^2$ which are the units of analysis.

If the researcher should have good evidence that the distributions sampled are normal, then he could use Bartlett's test or Hartley's test in the case of a single factor study, or perhaps, Overall and Woodward's test for a multifactor study. The first two tests are considerably more powerful than Box's test. The researcher must realize that these tests are extremely sensitive to violation of the normality assumption and should not be used unless the case for normality is strong.

Other tests could be recommended for specific distributions, but the researcher seldom, if ever, has knowledge of the type of distributions his samples have been drawn from. He could plot his data or calculate the first four moments to get an idea of the type of distribution he might be working with, but this will only give him a vague idea. Therefore, in most situations he should use a test which is relatively distribution free.

Recent literature suggests the use of Levene's test but in this study and others Levene's test has been shown to be too liberal for small sample sizes irrespective of the distribution sampled. The results of this study suggest that the Normal Scores extension of Levene's test would be a more reasonable test to consider. However, even this test is very poor should the populations sampled be exponential in nature.

In the mini-study used as a follow-up to the main study, the median deviated Levene type test showed considerable promise as being a valid contender. It is the only small sample test which appears to be well behaved at $\alpha = .10$ and $\alpha = .05$ for heavily skewed and leptokurtic distributions. The only distributions considered in this study were the exponential and the normal distributions in the case of the three-sample design with six observations per sample.

Shortly after the completion of this study, an article by Brown and Forsythe (6) provided more evidence to support the case of a median-deviated Levene type statistic. They considered balanced and unbalanced two-sample designs with sample sizes of ten, twenty or forty. They sampled from normal, Student's t on 4 degrees of freedom and chi-square on 4 degrees of freedom distributions. For all situations considered, the empirical estimates of α were very good and usually slightly conservative. Although the power of this test was somewhat lower than that of the variance ratio F test, it was competitive with it.

The use of median-deviated scores is supportable not only on the basis of these empirical studies but also on the basis of Miller's (42) analytic work. Miller has proven that Levene's test is asymptotically distribution free if the deviations are taken from the sample medians. This indicates that as $n \to \infty$, the null distribution of the test statistic will be invariant no matter what the distributions sampled look like. Whether this property holds for small sample sizes is a question for further empirical investigation.

The median-deviated Levene type test appears to be more promising than Box's test. Brown and Forsythe have supplied some evidence that the test is well behaved for unbalanced designs. In Gartside's (15) study, Box's test was shown to be considerably liberal if the design used was not a balanced one.

In light of the encouraging findings in the mini-study, the study of Brown and Forsythe and the analytic work of Miller, it is critical that the properties of this test be further explored. The test should be examined with a series of designs for a wide range of distributions with differing values of skewness and kurtosis.

Although the median-deviated statistic appears promising, other tests of better quality may emerge. Perhaps the use of a 25 percent trimmed mean would provide better estimates of type I error rate or better power than the use of the median. The 25 percent trimmed mean is the mean of the observations remaining after deleting the 25 percent largest and the 25 percent smallest values in that sample. The median could be considered a 50 percent trimmed mean. What trimmed mean would be the best is speculative.

The effect extreme values have on the sample mean seems to be the reason Levene's test is too liberal. Perhaps some transformation of the absolute deviations might reduce the effect of these extreme scores. Levene attempted log z and $\sqrt{z}$ but found both to give even more liberal estimates of type I error rate than the original z's. A transformation which might work wonders with one distribution could well be counter productive with another. It seems unlikely that there is a transformation which would work well for a wide variety of distributions.

The Jackknife technique proposed by Miller (42) seems to give somewhat liberal estimates for most distributions. This again could be due to the nonrobust quality of the arithmetic mean as an estimator of central location. Perhaps the Jackknife principle could be extended with a different estimator of central location. If the median were used, this would involve an ANOVA on v_{ij} scores where for observation j in sample i:

(1) $v_{ij} = n_i \log t_i^2 - (n_i - 1) \log t_{i(j)}^2$

(2) where $t_i^2 = \frac{1}{n_i-1} \Sigma_j (x_{ij}-m_i)^2$

(3) with m_i the median for all n_i observations

(4) and $t_{i(j)}^2 = \frac{1}{n_i-2} \underset{K\neq j}{\Sigma} (x_{ij} - m_{i(j)})^2$

(5) with $m_{i(j)}$ the median for the $(n_i - 1)$ observations remaining when the K*th* observation is removed

Without the use of a computer, this technique would be unbearably tedious to consider. Even with a computer, the large amount of ranking involved would make this a relatively expensive statistic to compute.

The other hope for a good test of variance homogeneity might be an extension of one of the two sample nonparametric tests of scale. Of those available, Mood's test has better asymptotic relative efficiency than either the Siegel-Tukey test or the Freund-Ansari test (5). However these three tests have disadvantages in the two-sample case. No exact tables have been prepared for Mood's test so it is referred to the normal distribution, thus the test is probably not too accurate if n is small. But the test fails on more serious grounds than this. For example, if the two-sample sizes are equal and the two populations do not overlap, the test is certain to accept the hypothesis of equal variance whether or not it is really true. It has been suggested that this might be corrected by aligning the sample medians, but Bradley (5, p. 120) points out that even "if the influence of unequal locations can be completely and satisfactorily eliminated, it is easy to invent populations having identical medians and identical dispersion indices of a given type but having shapes whose differences would cause the test to reject." Unfortunately this is also true with the other two nonparametric tests mentioned.

SELECTED BIBLIOGRAPHY

SELECTED BIBLIOGRAPHY

1. Bartlett, M. S. "Properties of Sufficiency and Statistical Tests." *Proceedings of the Royal Society of London*, Series A, 160 (1937): 268-82.

2. ________, and Kendall, D. G. "The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation." *Journal of the Royal Statistical Society*, Series B, 8 (1946): 128-38.

3. Box, G. E. P. "Non-normality and Tests on Variances." *Biometrika*, 40 (1953): 318-35.

4. ________, and Andersen, S. L. "Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumption." *Journal of the Royal Statistical Society*, Series B, 17 (1955): 1-26.

5. Bradley, J. V. *Distribution-Free Statistical Tests*. Englewood Cliffs, N.J.: Prentice Hall, Inc., 1968.

6. Brown, M. B., and Forsythe, A. B. "Robust Tests for the Equality of Variances." *Journal of the American Statistical Association*, 69 (1974): 364-67.

7. Cadwell, J. H. "Approximating to the Distributions of Measures of Dispersion by a Power of χ^2." *Biometrika*, 40 (1953): 336-46.

8. Conover, W. J. *Practical Nonparametric Statistics*. New York: John Wiley and Sons, Inc., 1971.

9. Coveyou, R. R. "Serial Correlation in the Generation of Pseudo-Random Numbers." *Journal for the Association of Computing Machinery*, 7 (1960): 72-74.

10. Feder, P. I. "On the Nonrobustness of the Jackknife and Box-Andersen Procedures for Estimating Variances in Small Samples." Paper presented at the Annual Meeting of the American Statistical Association, New York, December, 1973.

11. Fisher, R. A., and Yates, F. Statistical Tables for Biological, Agricultural and Medical Research. New York: Hafner Publishing Company, Inc., 1948.

12. Foster, L. A. "Testing for Equality of Variances." Dissertation Abstracts, 26 (1965): 1060B.

13. Gabriel, K. R., and Lachenbruch, P. A. "Non-Parametric ANOVA in Small Samples: A Monte Carlo Study of the Adequacy of the Asymptotic Approximation." Biometrics, 25 (1969): 593-96.

14. Games, P. A., Winkler, H. B., and Probert, D. A. "Robust Tests for Homogeneity of Variance." Educational and Psychological Measurement, 32 (1972): 887-910.

15. Gartside, P. S. "A Study of Methods for Comparing Several Variances." Journal of the American Statistical Association, 67 (1972): 342-46.

16. Gehan, E. A., and Thomas, D. C. "The Performance of Some Two Sample Tests in Small Samples With and Without Censoring." Biometrika, 56 (1969): 127-32.

17. Glass, G. V. "Testing Homogeneity of Variances." American Educational Research Journal, 3 (1966): 187-90.

18. ________, Peckham, P. D., and Sanders, J. R. "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analysis of Variance and Covariance." Review of Educational Research, 42 (1972): 237-88.

19. ________, and Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

20. Glendening, L. Personal Communication.

21. Greenberger, M. "An a Priori Determination of Serial Correlation in Computer Generated Random Numbers." *Mathematics of Computations*, 15 (1961): 383-89.

22. Hall, I. J. "Some Comparisons of Tests for Equality of Variances." *Journal of Statistical Computation and Simulation*, 1 (1972): 183-94.

23. Han, C. "Testing the Homogeneity of Variances in a Two-Way Classification." *Biometrics*, 25 (1969): 153-58.

24. Hartley, H. O. "The Maximum F-Ratio as a Short Cut Test for Heterogeneity of Variance." *Biometrika*, 37 (1950): 308-12.

25. Hastings, C., Jr. *Approximations for Digital Computers*. Princeton, N.J.: Princeton University Press, 1955.

26. Hector, M. A. "Evaluation of an Instructional Model for Teaching Counselor Trainees How to Establish Behavioral Objectives in Counseling." Ph.D. dissertation, Michigan State University, 1973.

27. Hodges, J. L., Jr., and Lehmann, E. L. "Comparison of the Normal Scores and Wilcoxon Test." *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability*, Berkeley and Los Angeles: University of California Press, 1961.

28. Holloway, C., and Capp, W. B. *F-Distribution Generator: A Fortran IV Program*. 1959. Revised by McKelvey, R. J., 1961.

29. Hull, T. E., and Dobell, A. R. "Random Number Generators." *SIAM Review*, 4 (1962): 230-54.

30. ________, and ________. "Mixed Congruential Random Number Generators for Binary Machines." *Journal for the Association of Computing Machinery*, 11 (1964): 31-40.

31. Kirk, R. *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, Ca.: Wadsworth, 1968.

32. Klotz, J. "Small Sample Power and Efficiency for the One Sample Wilcoxon and Normal Scores Tests." *Annals of Mathematical Statistics*, 34 (1963): 624-32.

33. Kruskal, W. H., and Wallis, W. A. "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association*, 47 (1952): 583-621.

34. Layard, M. W. J. "Robust Large Sample Tests for Homogeneity of Variances." *Journal of the American Statistical Association*, 68 (1973): 195-98.

35. Leaverton, P., and Busch, J. J. "Small Sample Power Curves for the Two Sample Location Problem." *Technometrics*, 11 (1969): 229-307.

36. Levene, H. "Robust Tests for Equality of Variances." *Contributions to Probability and Statistics.* Edited by I. Olkin et al. Stanford, Ca.: Stanford University Press, 1960.

37. MacLaren, M. D., and Marsaglia, G. "Uniform Random Number Generators." *Journal for the Association of Computing Machinery*, 12 (1965): 83-89.

38. McSweeney, M. T. "An Empirical Study of Two Proposed Nonparametric Tests for Main Effects and Interaction." *Dissertation Abstracts*, 28 (1968): 4005A-4006A.

39. ________, and Penfield, D. "The Normal Scores Test for the c-Sample Problem." *The British Journal of Mathematical and Statistical Psychology*, 22 (1969): 177-92.

40. Merrington, M., and Thompson, C. M. "Tables of Percentage Points of the Inverted Beta (F) Distribution." *Biometrika*, 33 (1943): 73-88.

41. Miller, A. J. Letter to the Editor. *Technometrics*, 14 (1972): 507.

42. Miller, R. G., Jr. "Jackknifing Variances." *Annals of Mathematical Statistics*, 39 (1968): 567-82.

43. Mood, A. M. "On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests." *Annals of Mathematical Statistics*, 25 (1954): 514-22.

44. Moses, L. E. "Rank Tests of Dispersion." *Annals of Mathematical Statistics*, 34 (1963): 973-83.

45. Naylor, T. H., Balintfy, J. L., Burdick, D. S., and Chu, K. Computer Simulation Techniques. New York: John Wiley and Sons, Inc., 1966.

46. Neave, H. R., and Granger, C. W. J. "A Monte Carlo Study Comparing Various Two Sample Tests for Differences in Mean." Technometrics, 10 (1968): 509-22.

47. Neel, J. H., and Stallings, W. M. "A Monte Carlo Study of Levene's Test of Homogeneity of Variance: Empirical Frequencies of Type I Error in Normal Distributions." Paper presented at the American Educational Research Association Convention, Chicago, April, 1974.

48. Neyman, J., and Pearson, E. S. "On the Problem of K Samples." Bulletin de l'Academie Polonaise des Sciences et des Lettres, June 1931, pp. 460-81.

49. Overall, J. E., and Woodward, J. A. "A Simple Test for Heterogeneity of Variance in Complex Factorial Designs." Psychometrika, 39 (1974): 311-18.

50. Pratt, J. W. "Robustness of Some Procedures for the Two-Sample Location Problem." Journal of the American Statistical Association, 59 (1964): 665-80.

51. Scheffé, H. The Analysis of Variance. New York: John Wiley and Sons, Inc., 1959.

52. Shorack, G. R. "Nonparametric Tests and Estimation of Scale in the 2-Sample Problem." Dissertation Abstracts, 26 (1966): 6751B.

53. Shukla, G. K. "An Invariant Test for the Homogeneity of Variances in a Two-Way Classification." Biometrics, 28 (1972): 1063-72.

54. Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Company, Inc., 1956.

55. ________, and Tukey, J. W. "A Nonparametric Sum of the Ranks Procedure for Relative Spread in Unpaired Samples." Journal of the American Statistical Association, 55 (1957): 429-45.

56. Sytsma, S. Personal Communication.

57. Terry, M. E. "Some Rank Order Tests Which Are Most Powerful Against Specific Parametric Alternatives." Annals of Mathematical Statistics, 23 (1952): 346-66.

58. Van Der Laan, P. "Exact Power of Some Rank Tests." Publication de l'Institut de Statistique de l'Université de Paris, 13 (1964): 211-34.

59. ________, and Oosterhoff, J. "Monte Carlo Estimation of the Powers of the Distribution-Free Two Sample Tests of Wilcoxon, Vander Waerden and Terry and Comparison of These Powers." Statistica Neerlandica, 19 (1965): 265-75.

60. ________, and ________. "Experimental Determination of the Power Functions of the Two Sample Rank Tests of Wilcoxon, Vander Waerden and Terry by Monte Carlo Techniques-I." Normal Parent Distributions. Statistica Neerlandica, 21 (1967): 55-68.

61. Weber, J. M. "The Heuristic Explication of a Large-Sample Normal Scores Test for Interaction." British Journal of Mathematical and Statistical Psychology, 25 (1972): 246-56.

62. Wheeler, D. J. "An Alternative to an F Test on Variances." Dissertation Abstracts, 31 (1971): 6334B.

63. Winslow, S. S., and Arnold, J. C. "A Stochastic Simulation Study of a Rank-Like Test for Dispersion Which Is Distribution-Free Under Location Differences." Journal of Statistical Computation and Simulation, 1 (1972): 315-29.